Eliciting Private Information from Selfish Agents

Rafael Frongillo

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2013-138
http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-138.html

August 8, 2013

Copyright © 2013, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Acknowledgement

The author was supported by NSF grant CC-0964033, a Google University Research Award, and the National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

Eliciting Private Information from Selfish Agents

by

Rafael M. Frongillo

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Christos Papadimitriou, Chair
Professor Peter Bartlett
Associate Professor David Ahn

Fall 2013

Eliciting Private Information from Selfish Agents

Copyright 2013
by
Rafael M. Frongillo

Abstract

Eliciting Private Information from Selfish Agents

by

Rafael M. Frongillo

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Christos Papadimitriou, Chair

Ever since the Internet opened the floodgates to millions of users, each looking after their own interests, modern algorithm design has needed to be increasingly robust to strategic manipulation. Often, it is the inputs to these algorithms which are provided by strategic agents, who may lie for their own benefit, necessitating the design of algorithms which incentivize the truthful revelation of private information – but how can this be done? This is a fundamental question with answers from many disciplines, from mechanism design to scoring rules and prediction markets. Each domain has its own model with its own assumptions, yet all seek schemes to gather private information from selfish agents, by leveraging incentives. Together, we refer to such models as *elicitation*.

This dissertation unifies and advances the theory of incentivized information elicitation. Using tools from convex analysis, we introduce a new model of elicitation with a matching characterization theorem which together encompass mechanism design, scoring rules, prediction markets, and other models. This lays a firm foundation on which the rest of the dissertation is built.

It is natural to consider a setting where agents report some alternate representation of their private information, called a *property*, rather than reporting it directly. We extend our model and characterization to this setting, revealing even deeper ties to convex analysis and convex duality, and we use these connections to prove new results for linear, smooth nonlinear, and finite-valued properties. Exploring the linear case further, we show a new four-fold equivalence between scoring rules, prediction markets, Bregman divergences, and generalized exponential families.

Applied to mechanism design, our framework offers a new perspective. By focusing on the (convex) consumer surplus function, we simplify a number of existing results, from the classic revenue equivalence theorem, to more recent characterizations of mechanism implementability.

Finally, we follow a line of research on the interpretation of prediction markets, relating a new stochastic framework to the classic Walrasian equilibrium and to stochastic mirror descent, thereby strengthening ties between prediction markets and machine learning.

To Jenn

♡

Contents

List of Figures

List of Tables

Acknowledgments

I am deeply indebted to Christos for his patience and persistence in advising me, and for believing in me even when I did not. His advice, always candid and poetic, has guided me through many tough decisions. But even more so, I am continually inspired by his incredible enthusiasm and curiosity, and his courage to follow it wherever it leads him.

I have been fortunate to have many unofficial advisors as well. I am grateful to my "deputy advisor" Jake, for taking me under his wing, for all the support and tough love, and for constantly reminding me that I couldn't have done this without him. I also thank Sarah Day for her advice, support, and friendship, even when I disappeared for months at a time. Looking further back, I still owe thanks to those who helped me get to Berkeley in the first place, and whose teaching and advice still run through my veins: John Smillie, Dave Bock, Roselyn Teukolsky, and Mr. Demo.

It takes a university to raise a doctorate, and I cannot think of a better place to be raised than UC Berkeley. The vibrant, warm, and exceptionally intelligent computer science theory group at Berkeley has made my time here productive and fun. I'm thankful to James for inspiring me to take more walks; to Anindya and Isabelle for being the best officemates one could ask for; to Greg for squeezing my head in the orange room, and all the advice over the years; to Chris, George, and Yaron for all the conversations and for keeping EconCS alive and well; to Piyush for all of the technical (and mathematical) support; to Ron Lavi and Chris for the very enlightening discussions on mechanism design; to Guoming, Tom, Siu {On|Man}, Andrew, Ma'ayan, Anand, Antonio, Di, Woo, Urmila, Jonah, Andrew, Paul, Alex, Aviad, Tselil, Rishi, Jarett, Jonah, Ben, and the rest of the current theory crew; to Omid, Alexandra, Henry, Lorenzo, Meromit, Ilias, Madhur, Grant, Paul, Thomas, Alexandre, Virginia, and all the other recent Berkeley graduates with whom I've had the pleasure of meeting and working with over these years. Beyond Soda Hall, I would like to thank all of the singers, soccer players (especially the Greeks!), Capoeiristas, ultimate players, hackers (i.e. Adam), and co-opers for keeping me grounded by reminding me that there is life beyond the whiteboard.

Some of my most memorable and productive experiences have been away from the Berkeley campus. I thank Éva Tardos for hosting me at Cornell for a summer, despite the setbacks. I'm grateful to Bob and Mark at NICTA for their generous hospitality in Canberra, and to Nikete for his own flavor of the same. Thanks to Yoram and Indraneel for an awesome Google internship, and to Kevin, Peter, and Ryan for topping it off with some good chords. While we met but once, I am grateful to Nicolas Lambert for the many insightful email conversations, and for inspiring me to work on many of the problems in this thesis. I also am grateful to Miro, Sébastien, Jenn, and other MSR-NYC researchers for their input. Finally, I cannot understate the importance of my experience at Microsoft Research Cambridge (UK) to this dissertation; I owe a great deal of thanks to Ian for being so generous with his time and creating the collaborative environment from which much of this thesis grew.

Lastly, I am *profoundly* grateful for the unyielding and unconditional love and support of my close friends and family. I cannot begin to express my appreciation for you all.

Chapter 1

Introduction

Information is the currency of our age. Entire occupations and disciplines are devoted to creating, distributing, and exchanging information — and this information economy, run on the platform of the internet, has become the nervous system of our civilization. Of course, this nervous system is comprised of many different agents, often algorithmic, and it is a natural question to ask how this information is, and should be, traded and exchanged in this strategic environment. In other words, if information is a currency, how should we design the currency exchange?

The field of algorithmic economics, also called algorithmic game theory, has a natural vantage point to answer this question. This dissertation embraces this algorithmic and economic perspective, and focuses on the one-sided exchange of information, that is, the exchange of privately-held information for another currency or commodity.

Unlike most currencies, however, information is not always verifiable. Your bank will confirm the amount of money received, but how would you know how accurate a weather forecast is? It is easy to see that without any external signals, accuracy of information is entirely subjective. To regain objectivity, we will make use of the fact that agents are *strategic*, and will behave in such a way as to maximize their own welfare. This basic economic model of behavior will let us leverage the *incentives* of the agent, to ensure honest reports. This is what we mean by *elicitation* — the study of mechanisms to incentivize the truthful reporting of private information by selfish agents.

This dissertation studies elicitation from a theoretical perspective, with an eye toward characterization. That is, we seek to determine exactly when and how elicitation schemes work. Our study spans many different models and settings, but for the most part we focus our attention on three core models: scoring rules, prediction markets, and mechanism design.

By and large, we will find that a unified study of an elicitation is not only possible, but fruitful. Armed with tools from convex analysis, our main result is a general model and characterization theorem which covers all three of these core models. From there, we explore the intricacies of each domain, focusing on more nuanced characterizations and highlighting deep connections among the models and to machine learning, geometry, statistics, and other domains.

After some nontechnical motivation in §1.1, we introduce the three core models in §1.2. We will motivate our focus on characterizations in §1.3, followed by a more thorough overview of our results in §1.4.

1.1 A tale of two elicitations

We begin with a pair of whimsical parables to motivate our study, hoping to convey both the diversity and importance of elicitation models. The strictly technically-minded reader may wish to skip straight to the literature review in §1.2.

1.1.1 When it rains, it ~~pours~~ doesn't rain

Meteorology is hardly a dry subject in the state of New York. After a successful career leading a weather station in Ithaca, NY, our protagonist Marguerite found herself overseeing all 23 weather stations in the state. Her primary responsibility was now to ensure the accuracy of weather reports throughout the state. Being particularly interested in precipitation, she instructed all stations to forecast, among their other predictions, the likelihood of rain $p \in [0, 1]$ and the amount of rainfall $r \in \mathbb{R}_+$ in inches.[1]

After a few weeks, Marguerite took a look at forecasts given under her watch thus far, comparing them to what weather was actually recorded. To her dismay, while some stationmasters appeared to be doing their job, others were clearly not putting in as much effort to make accurate predictions, some reporting apparently at random, and others issuing the same forecast each day despite wildly changing conditions. And after some thought, she realized, why *would* the stationmasters put in the effort? They were being paid to issue forecasts, but they had no skin in the game; their accuracy was not being enforced or even evaluated in any way.

Well, no more. Marguerite had a brilliant idea: make the salaries of the stationmasters depend on the accuracy of their reports! Excited to try out her new plan, she informed all 23 stationmasters that 50% of their salary would be fixed, but the other 50% would be a function of how accurate their reports were, particularly with regard to precipitation. As soon as she had made the announcement, however, Marguerite realized that she needed to come up with a way to *quantify* the accuracy. Not thinking much of it, and without announcing her choice, she settled on a simple daily score: \$50 times the reported probability of what actually happened (i.e. $\$50 \cdot p$ if it rains and $\$50 \cdot (1 - p)$ otherwise), plus a bonus \$30 if the rainfall r was within 0.1 inches of being correct. More precisely, the score would be

$$\$50 \cdot (p\, \mathbb{1}\{\text{rain}\} + (1 - p)\mathbb{1}\{\text{no rain}\}) + \$30 \cdot \mathbb{1}\{|r - r^*| < 0.1\},$$

where $\mathbb{1}$ is the indicator, and r^* is the actual rainfall value.

At first, the plan worked phenominally. Reports were more accurate than ever, and everyone seemed to be putting much more energy and effort into their predictions. When payday

[1] We will say "rain" to refer to any type of precipitation.

arrived, however, several stationmasters were enraged by their salaries, and demanded to know exactly how Marguerite was calculating the accuracy. She obliged and announced her formula to all, quelling the upset.

Unfortunately for Marguerite, the next month did not bode as well. After receiving numerous complaints from citizens about the weather reporting, she took a look at the forecast history once again. What she found surprised her. Not only were the reports bad, they were substantially worse than before. And even the top meteorologists were botching their forecasts. After examining the data, she noticed some peculiar patterns. The first was that, except for the first few days after her announcement, nearly *all* probabilistic forecasts were either 0 or 1 — that is, everyone was reporting with complete certainty whether or not it would rain. The second pattern was even more shocking: almost a third of the "100% chance of rain" forecasts were accompanied by a projection of *0 inches of rainfall*.

After discussing with colleagues, who were equally confused, Marguerite decided it must be her scoring mechanism. She did a little research and discovered that a more common way to evaluate rainfall forecasts was to use the *relative error*, or $\text{RE}(r, r^*) = |(r - r^*)/r|$. Satisfied, she announced the change of the rainfall bonus from $\$30 \cdot \mathbb{1}\{|r - r^*| < 0.1\}$ to $\$30 \cdot (1 - \text{RE}(r, r^*))$, thus rewarding smaller relative errors. (Note that this may yield a negative rainfall bonus if the reported rainfall is orders of magnitude too low, in which case stationmasters would *lose* money.) To her relief, the rainfall estimates improved, but a few weeks later she began getting calls about even more preposterous forecasts: stations were reporting a 0% chance of rain with almost 2 inches of rainfall!

Exasperated, Marguerite was determined to solve the mystery of these bogus forecasts. Comparing the data from each of the three months, she finally figured out what was going on. Starting with the probabilities, she put herself in the stationmasters' shoes and imagined that she thought rain was 20% likely. Given that she would be paid $\$50 \cdot p$ if rain occured, and $\$50 \cdot (1 - p)$ otherwise, it is clear that on average she is better off forgetting about the rain and reporting $p = 0$; this gives an average of \$40, which certainly beats the average of $\$50 \cdot ((0.2)^2 + (0.8)^2) = \$50 \cdot 0.68 = \$34$ for reporting her actual belief. In general, Marguerite's "linear score" was incentivizing the stationmasters to put all weight on one of the two extremes.

Similar thinking told her that the first rainfall bonus $\$30 \cdot \mathbb{1}\{|r - r^*| < 0.1\}$ would be optimized by reporting the *mode*[2] of one's rainfall distribution, while the second bonus $\$30 \cdot (1 - \text{RE}(r, r^*))$ would yield a skewed median, favoring much higher values. This explains why forecasters would report such bizarre combinations of forecasts: if one's belief about the rainfall were highly uncertain, the mode could be 0 while the total mass of $r > 0$ could be above 0.5, and the first bonus would yield forecast "rain but no rainfall"; if a heavy storm were possible but unlikely, the second bonus would incentivize the report "heavy rainfall but no rain." See Figure 1.1 for an illustration.

Glad to have solved the mystery, Marguerite used her line of reasoning to search for a better scoring rule. How could she design a score whose incentives led to good forecasting?

[2]The mode of a distribution is the value with highest probability density.

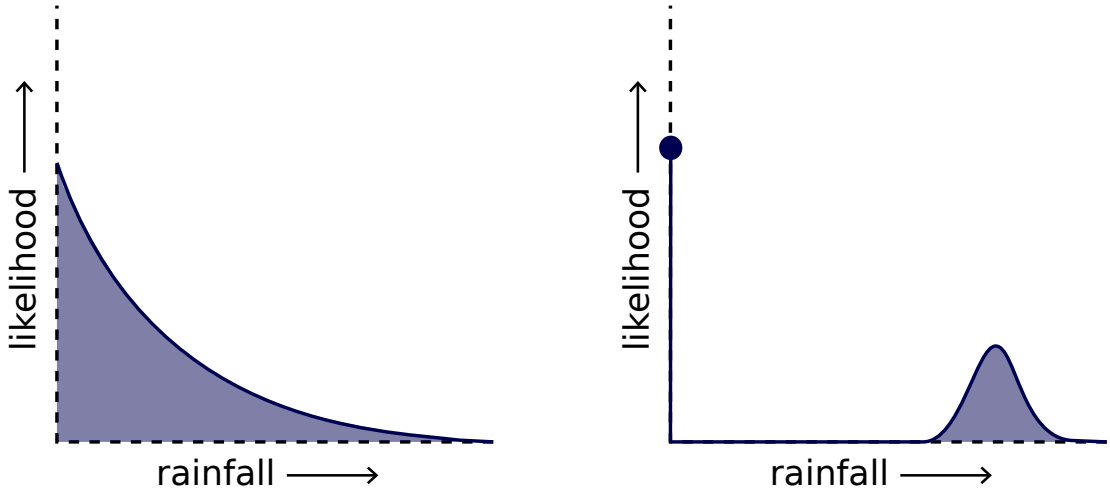


Figure 1.1: Two beliefs about rainfall values, represented here as probability measures (the dot in (b) is a point mass on the value 0). Under the first rainfall bonus, belief (a) would yield the report "rain but no rainfall," while under the second rainfall bonus, belief (b) would yield "rainfall but no rain."

It was then that she remembered something her old friend Carl Friedrich Gauss taught her: the minimizer of $(x-y)^2$ on average over x is simply y. Inspired, she changed *both* scores accordingly, yielding the new daily score of $\$100-\$50\cdot(p-\mathbb{1}\{\text{rain}\})^2-\$30(r-r^*)^2$, and as she easily verified, the score is maximized by forecasting one's true belief about the occurrence of rain, and one's *mean* belief about the rainfall. And thus it was that rain forecasts for New York became so impeccably accurate.

1.1.2 Of rice and men

Ed was never terribly prone to charity, but he also hated rice. So when he accidentally entered, and won, a raffle for lifetime monthly shipments of 40 bushels of rice, he decided it was time to do a little good for the world. He happened to live a 20-minute drive from a poor village. His calling was clear: each month, when the bushels arrived, go to the village and distribute the rice.

April. After a warm greeting by the villagers, Ed explained that he wanted the rice to go to those who were struggling most, and proceeded to ask each family how much rice they would need for the next month. Tallying all of the numbers, he found the total demand from the village to be 80 bushels, double his supply. To be equitable, he halved each request, distributed the rice accordingly, and returned home.

May. To be sure that he was continuing to serve the needs of the villagers, Ed once again asked for each family's needs for the month of May. To his surprise, everyone was in greater need, so that in total the village needed 160 bushels. Ed nevertheless distributed the rice proportionally, feeling slightly concerned that his rice would not go as far. Upon leaving he announced that next time, due to his limited supply, no single family could request more than a bushel.

June. When every family requested a bushel, Ed was hardly shocked, but what did surprise him was how many new families had emerged — and no family was larger than 2 people! Surely familial schisms on such a massive scale could not be coincidental. After again distributing his rice proportionally, indeed completely evenly this time, Ed started thinking a bit more about what had transpired.

Clearly, he had failed to recognize that the families might be *strategizing* instead of blindly telling him the truth. In May, villagers must have realized that his proportional distribution meant that in order to get their request, they had to *overshoot* their needs. In June, they got even more creative and divided their families up so that each "real" family would have a higher relative share. Despite Ed's efforts to distribute the rice according to the needs of the villagers, the data he was getting from them were essentially meaningless.

Ed needed a way to make honesty in the best interests of the villagers, but how? Then it dawned on him that instead of fighting their clearly excellent ability to strategize, he could use it to his advantage. Inspired, he devised a new plan.

July. Armed with a calculator and feeling clever, Ed returned to the village and announced his scheme. Each family would tell him 10 numbers, their value in dollars for each additional fifth of a bushel. Then, he would give out the rice to maximize the *total value* of the villagers, and additionally, he would *pay* each family an amount equal to what the other families valued their allotment. For example, if Ed had just 2 bushels and families A, B, and C reported values (in USD) of $v_A = [12, 12, 12, 12, 0, 0, 0, 0, 0, 0]$, $v_B = [4, 4, 4, 4, 4, 4, 4, 4, 4, 4]$, and $v_C = [34, 21, 13, 8, 5, 3, 2, 1, 1, 0]$, then the optimal allotment would be $4/5$ of a bushel to A, $1/5$ of a bushel to B, and 1 bushel to C; the families value these allotments respectively at \$48, \$4, and \$81, so the payments would then be $p_A = \$85$, $p_B = \$129$, and $p_C = \$52$. The villagers readily agreed to this plan, and thankfully they provided Ed with much more reasonable answers this time. Once all the numbers were collected, he set to work calculating the rice shares and corresponding payments.

At sunset, Ed realized that his calculator was not up for the task — to figure out the workloads, he needed to compute nearly a million intermediate values. Instead, he made up some plausible numbers for the rice, and then readily computed the payments. On his way home, being exhausted and \$1542 poorer himself, Ed vowed to simply drop off the rice next time and let these clever villagers figure out how to distribute it themselves.

1.1.3 Discussion

These parables show the diversity of elicitation models, and give a flavor of the challenges that arise in designing schemes to incentivize honesty. The first motivates the study of *scoring rules*, which we introduce in § 1.2.1, and illustrates how challenging it can be to assess forecasts; indeed, all scoring functions mentioned have actually been used in practice, and the relative error is still in use [50]. While some of the difficulty Marguerite faced came from an ill-specified notion of an "accurate report," the basic problem was the lack of a *calibrated* score — she wanted honest and accurate forecasts to give the maximum average score. We will see many variants of this model throughout the dissertation.

The second parable addresses a problem in *mechanism design*, a branch of microeconomic theory. Specifically, Ed was trying to implement a *multi-unit auction*, having multiple bushels of rice and wishing to "auction" them off to the villagers. His June attempt suffered from a lack of *sybil-proofness*, meaning that agents had an incentive to split their identity into several copies; we will not address such concerns in this dissertation, though in practice this is a very important consideration especially in computational settings. The scheme from July is the most sophisticated, and is called the VCG mechanism; see § 1.2.3.

Unfortunately, as Ed saw, his July scheme suffered from two problems: a high implementation cost, and computational intractability. The first problem is actually easy to fix; in addition to receiving cash equal to the value of the rice everyone else received, Ed could have had each family f provide labor (e.g. work in his orchard) a number of hours equivalent to the value of what everyone *would have gotten* had f not even participated to begin with. Returning to the example above, at a labor cost of \$10/hour, this would yield labor amounts of $L_A = (p_A + 16)/10 = 10.1$ hours, $L_B = (p_B + 3)/10 = 13.2$ hours, and $L_C = (p_C + 20)/10 = 7.2$ hours. This payment and rebate scheme is called the *Clarke pivot rule* (Clarke is the 'C' in 'VCG'). Note that as the amount of labor for family f does not depend on the reports of f, the scheme is still truthful. Moreover, Ed actually stands to gain in the transaction, as net payments to villagers $(p_f - 10L_f)$ are non-positive. For the computational intractability, Ed should have familiarized himself with recent approximation results in algorithmic mechanism design; for example, Dobzinski and Nisan [40] give an efficient mechanism which is truthful and which approximates the optimal rice distribution.

1.2 The literature on elicitation

1.2.1 Scoring rules

Many authors point to the paper of Glenn W. Brier [26] as the earliest mention of what we now call a *proper scoring rule*, or a *proper loss* in machine learning.[3] The paper, published in 1950 in the *Monthly Weather Review*, observed that verifying probability forecasts can

[3] In this dissertation, we will use gains rather than losses, but all results pertaining to scoring rules hold equivalently for proper losses as well.

be challenging, and proposed a tool, now known as the Brier score, to measure predictions once an outcome is known. It is instructive to quote directly from Brier, to get a sense of his approach:

> Suppose that on each of n occasions an event can occur in only one of r possible classes or categories and on one such occasion, i, the forecast probabilities are $f_{i1}, f_{i2}, \ldots, f_{ir}$, that the event will occur in classes $1, 2, \ldots, r$ respectively. The r classes are chosen to be mutually exclusive and exhaustive so that
>
> $$\sum_{j=1}^{r} f_{ij} = 1, \quad i = 1, 2, 3, \ldots, n$$
>
> A number of interesting observations can be made about a vertification score P defined by
>
> $$P = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n} (f_{ij} - E_{ij})^2 \tag{1.1}$$
>
> where E_{ij}, takes the value 1 or 0 according to whether the event occurred in class j or not.

One such interesting observation, as Brier points out, is that if p_j is the actual frequency of the event E_{ij} over i, then p_j will minimize the score P. (Of course, in the modern formulation agents seek to *maximize* their score, which just amounts to negating P.) This "calibration" property, that the optimal forecast is equal to the actual frequency, is what we mean when we say a scoring rule is *proper*.

In its simplest incarnation, a scoring rule is simply a function $\mathfrak{S}(\cdot, \cdot)$ taking two inputs, a probability forecast $\hat{p}$ and an outcome x from some set of outcomes. The scoring rule is *proper* if it is calibrated in the sense that if x is sampled according to some "true" distribution p then for any $\hat{p}$

$$\mathbb{E}_p[\mathfrak{S}(p, x)] \geq \mathbb{E}_p[\mathfrak{S}(\hat{p}, x)].$$

In other words, the forecaster maximizes the expected value of $\mathfrak{S}$ by reporting the true distribution.[4]

One of the most well-known examples is the *logarithmic scoring rule* defined by $\mathfrak{S}(p, x) := \log p(x)$, which was introduced in 1952 by I.J. Good in an independent study [52]. One can check that once again the logarithmic score is proper, in that the expected score is maximized when reporting the true distribution p.

[4]Interestingly, while propriety is a very natural condition for a scoring rule to have, and considered necessary by some (cf. Gneiting [50]), it is quite common even today to encounter uncalibrated forecasts in statistics, econometrics, meteorology, machine learning, and many other disciplines.

Already in 1956, John McCarthy [71] observed the crucial role of convexity in this theory of forecast scoring, noting that in some sense a scoring rule must be a derivative of a convex function. In fact, the convex function is given by $f(p) = \mathfrak{S}(p, p)$. He gives intuition for this phenomenon as follows:

> The intuitive content of the convexity restriction is that it is always a good idea to look at the outcome of an experiment if it is free. For suppose that the experiment has two outcomes, A and A^*, which would give one probabilities p and p^* for the event in question. Let t be the probability that A is the outcome. If we decide not to look, our expectation is $f(tp + (1-t)p^*)$, while if we decide to look, our expectation is $tf(p)+(1-t)f(p^*)$.

In other words, the agent has a prior belief over distributions, with weight t on p and $1 - t$ on p^*, so before the experiment, the belief about the outcome is simply $\hat{p} = tp + (1 - t)p^*$. Not running the experiment gives an expected payout of $t\mathfrak{S}(\hat{p}, p) + (1-t)\mathfrak{S}(\hat{p}, p^*) = \mathfrak{S}(\hat{p}, \hat{p})$, which our definition of $f(\hat{p})$. If the agent decides to carry out the experiment, however, she will know the distribution of the outcome with certainty; thus the expected post-experiment score, as computed *before* the experiment, is $t\mathfrak{S}(p, p) + (1 - t)\mathfrak{S}(p^*, p^*)$. Thus, if we want the agent to have an incentive to perform this free experiment (i.e. to use all available information), we must have $f(\hat{p}) \geq tf(p) + (1 - t)f(p^*)$, meaning f must be a convex function.

Building on McCarthy's paper, in 1971 Leonard Savage published a very general treatise on eliciting "personal probabilities," i.e. subjective probability estimates, in what is widely considered to be the seminal work on proper scoring rules [92]. He motivates his study by the problem of eliciting an agent's value for a commodity:

> Suppose the experimenter offers, once and for all, to buy some of a commodity at each possible price—more accurately price rate—so much at each rate. The subject will then have an incentive to satisfy the expressed demands of the experimenter at all rates higher than the subject's rate r but not those at lower rates, thereby revealing r.

That is, the experimenter proposes buying 1 unit of the commodity from the subject at price \$1000, another unit at \$999, a third at \$998, and so on until the reported value r is reached. Clearly, the transaction is profitable to the subject while the rate is above her true rate r, breaks even at r, and comes at a loss when the rate is below r; hence, the subject will report truthfully. We will revisit this clever idea in §3.4.3.

Applying the above result to the problem of eliciting probabilities, and more generally expectations of random variables, Savage argues (more explicitly than McCarthy, though still not completely rigorously) that a proper scoring rule should have the form,

$$\mathfrak{S}(p, x) = J(p) - J'(p) \cdot (p - \mathbb{1}_x), \tag{1.2}$$

where $J(p) = \mathfrak{S}(p, p)$ is any strictly convex function and $\mathbb{1}_x$ is the indicator vector for x. Savage goes on to give similar arguments that scoring rules for eliciting expectations should have the same form (1.2).

Since Savage's 1971 paper, there has been a vast and diverse array of research on the subject of proper scoring rules and related or extended models. A recent line of research focuses on eliciting statistics, or *properties*, of distributions; see § 3.1 and § 4.1 for more background. On the general topic of scoring rules, Gneiting and Raftery [51] give a more modern discussion, and extend and generalize many scoring rule characterizations in the literature, concluding that a generalizaton of the form (1.2) is necessary and sufficient even when working with probability measures on arbitrary measure spaces.

1.2.2 Prediction markets

In 1907, a contest was held at the annual show of the West of England Fat Stock and Poultry Exhibition in Plymouth, to judge the weight of an ox. Each of roughly 800 participants submitted an estimate, with the closest guesses winning prizes. After the fact, statistician Francis Galton took a closer look at the estimates, and found to his surprise that while each individual estimate was off by an average of roughly 3%, the *median* estimate was accurate within 0.8% [48]. While Galton viewed this as support of democracy as a sensible governing strategy, the general phenomenon that simple aggregations of estimates can yield very accurate predictions is now commonly known as the *wisdom of the crowd*.

Markets, and particularly speculative markets, can be seen as carrying out such "ox-weighing" contests every day. Traders buy and sell contracts directly linked to commodity, and are in essence predicting its future price. Motivated by this, several markets, called information markets or *prediction markets*, have arisen solely for the purpose of aggregating information. These markets allow traders to buy and sell contracts amongst themselves which are contingent on some future event or outcome. The contracts take a particular form, called an *Arrow-Debreu security*, which pays off \$1 if a given event occurs, and \$0 otherwise. A risk-neutral trader who thinks event A will happen with probability $p(A) = 0.3$ would therefore stand to gain by buying the A security for any price less than \$0.3, and by selling for any price greater. Hence, the market price in some sense reflects the "consensus belief" of the traders. (What precise sort of consensus belief one can glean from prediction market prices is a matter of debate; see Chapter 6.)

Thin markets and scoring rules

While prediction markets tend to produce highly accurate forecasts [19, 69, 101], they are not without problems. Aside from legal issues, prediction markets in this continuous double-auction format can suffer from a *thin market* problem. That is, trading volume can be very low, resulting in a wide buy-sell spread, i.e. a large discrepancy between the highest price at which one can sell and the lowest price at which one can buy. In the extreme, the market

may only have one trader who actually has information, and if everyone else is aware of their own ignorance, no trade will occur, and the market price will be meaningless.

To circumvent the thin market problem, Hanson [56] proposed designing a prediction market based on our old friend, the scoring rule. Instead of trading with each other, participants would place bets with a central authority, known as a *market maker*, which would continue to publish a joint forecast representing the "consensus hypothesis" of the market.

The framework itself is remarkably simple. The market maker publishes a proper scoring rule $\mathfrak{S}$ and an initial probability estimate p_0. On each round t in a sequence, the current consensus probability p_t is posted, and any trader can place a bet by *modifying* the probability to any desired value p_{t+1}. In the end, the true outcome x is revealed to the world, each trader receives a (potentially negative) profit of

$$\mathfrak{S}(p_{t+1}, x) - \mathfrak{S}(p_t, x). \tag{1.3}$$

Notice two facts about this framework: (a) if a trader at time t knows the true probability p^* then he always maximizes expected profit by setting $p_{t+1} = p^*$ and (b) because of the telescoping sum, if p_T is the final estimated probability then the market maker needs only to pay out a total of $\mathfrak{S}(p_0, x) - \mathfrak{S}(p_T, x)$. Hanson referred to this form of prediction market as a *market scoring rule* and, when the logarithmic scoring rule from above is used, this was called the Logarithmic Market Scoring Rule (LMSR).

Hanson's prediction market framework, which requires traders to make probability estimates and judges them according to a scoring rule, does not fit into our typical understanding of betting or financial markets, which as we described above have parties buy and sell contracts whose payoff is contingent on future outcomes. A natural question to ask is whether we can convert the market scoring rule betting language, in which traders are asked to directly report probability predictions, into one in which traders simply purchase Arrow-Debreu securities at prices set by the market maker.

Surprisingly, Hanson [56] showed that one could in fact view the LMSR as doing exactly this. Instead of asking for probabilities on events $1, 2, \ldots, n$, one can think of the market maker as asking for trade requests on corresponding securities. When the prices for these securities are adjusted very carefully in response to trades, the two settings would be equivalent:

> We can summarize all this by saying that each market scoring rule in essence has a "net sales so far" vector $s = \{s_i\}_i$, where each s_i says how many units have been sold of assets of the form "Pays \$1 if the state is i." The current unit price for a tiny amount of such an asset is p_i, and these prices change according to a price function $p(s)$, which is in essence a generalized inverse of the scoring rule function $s(p)$. For example, for the logarithmic scoring rule $s_i(p) = a_i + b\log(p_i)$, the price function is the exponential
>
> $$p_i(s) = \frac{e^{(s_i - a_i)/b}}{\sum_k e^{(s_k - a_k)/b}}. \tag{1.4}$$

Hence, in some sense, one can think of scores as being the securities themselves.

Is this just a cute trick using log and exp, or can this "securitization" be done for other market scoring rules? In 2007, Chen and Pennock [33] showed that such a security-based market is possible for a certain market scoring rules and proposed a market formulation based on a *cost function*; we briefly sketch their framework here. As in Hanson's reduction, some future outcome $i \in \{1, \dots, n\}$ will occur, and the market maker sells an Arrow-Debreu security for each outcome. However, in this setting, the market maker is endowed with a convex and differentiable cost function $C : \mathbb{R}^n \to \mathbb{R}$. The framework is as follows:

- Contract j pays \$1 if and only if outcome j occurs.
- The "quantity vector" $q^t \in \mathbb{R}^n$ is posted at each time t, where $q^t = 0 \in \mathbb{R}^n$.
- At any point in time, a trader may purchase a "bundle" of shares described by $r \in \mathbb{R}^n_{\geq 0}$; that is, r_i is the number of shares purchased for outcome i.
- The price for bundle r at time t is $C(q^t + r) - C(q^t)$.
- After selling r to the trader, the market maker updates $q^{t+1} \leftarrow q^t + r$.
- At the close of the market, some outcome i is revealed, and the market maker pays for all the winning contracts, a total cost of q_i^t.

Notice that the "current market price" is represented by the derivative $\nabla C(q)$, since $\nabla_i C(q)$ is the marginal cost of a tiny purchase of contract i. Thus, since the market prices in equilibrium are the expected return of the contract, $\nabla C(q)$ should be the market's probability vector. In fact, to avoid arbitrage opportunities, the market maker must ensure that $\nabla C(q)$ is always a distribution. Chen and Pennock [33] went on to show how to recover the LMSR using this cost-function framework. If we take

$$C(q) \doteq b \log \left(\sum_i \exp\{(q_i - a_i)/b\} \right), \tag{1.5}$$

for constants $b \in \mathbb{R}_+$, $a \in \mathbb{R}^n$, then one can check that

$$\nabla_i C(q) = \frac{e^{(q_i - a_i)/b}}{\sum_k e^{(q_k - a_k)/b}}, \tag{1.6}$$

thus recovering Hanson's price function (1.4).

Larger outcome spaces

An important problem with the prediction market frameworks we have described thus far is that they are not practical for large outcome spaces. Imagine a scenario where the outcome is a combinatorial object, like the joint outcome of a single-elimination tournament with

n teams. In the case of the market scoring rule, we must ask each participant to submit beliefs in the form of an *entire distribution* over the outcome space, here containing 2^{n-1} outcomes. In the cost-function framework, the market maker is required to sell an Arrow-Debreu security for each of these possible outcomes. Clearly neither of these will be feasible for large n. One natural solution is to consider a small set of marginal probabilities, and to have the betting language depend only on these values. It has been considered whether a market maker can efficiently simulate LMSR pricing within this betting language, yet a large number of these results have been negative [30, 29].

Abernethy et al. [1, 3] proposed a new framework for combinatorial prediction market design which avoids some of these hardness issues. The idea is best explained by way of example. Imagine a round-robin tournament which ends up with a (strict) ranking of all n teams. Rather than have a single contract corresponding to each of the $n!$ outcomes, a market maker can sell only $\binom{n}{2}$ contracts, one for each pair i, j corresponding to the predicate "does team i rank higher than team j?" This is often called a *complex* or *incomplete* market, as the traders can only express beliefs in this lower-dimensional contract space. Nevertheless, we can still use a cost function $C : \mathbb{R}^{\binom{n}{2}} \to \mathbb{R}$ to price these contracts as we did in the complete market setting. The market maker will maintain a quantity vector $q \in \mathbb{R}^{\binom{n}{2}}$, and will price a bundle of contracts $r \in \mathbb{R}^{\binom{n}{2}}$ according to the rule $C(q + r) - C(q)$. Given any final ranking of the n teams, we can describe the payoffs of all contracts by some $x \in \{0, 1\}^{\binom{n}{2}}$. The trader who previously purchased bundle r will receive $r \cdot x$.

In this setting, how ought we design C? Previously, in the complete market setting, we noted that ∇C should always be a distribution. Abernethy et al. [1] showed that, in a similar vein, C must have the property that $\{\nabla C(q) : q \in \mathbb{R}^{\binom{n}{2}}\}$ be the convex hull of all payout vectors x over all the $n!$ possible outcomes. Letting H denote this convex hull, they construct C via *conjugate duality* (see [89, 95] and §3.3). If R is some strictly convex function with domain H, then setting $C(q) \doteq \sup_{x \in H} x \cdot q - R(x)$ is sufficient to guarantee the desired properties of the market.

1.2.3 Mechanism design

The field of mechanism design is often considered the "engineering" side of economics. Instead of studying the economic properties and equilibria of an existing fixed system, here one considers *designing* a system to have certain properties or equilibria. In other words, in addition to the usual agents, we add an additional agent called the *mechansim designer* or *principal*, whose strategies are the payoff matrices that the other agents will face. Our tour of mechanism design will be brief, as our results pertaining to mechanism design in §2.3.1 and Chapter 5 are for the most part set in their historical context. Hence, we focus on a few key ideas that are central to the field and will come up throughout the dissertation.

Generally speaking, the mechanism designer chooses a *mechanism*, which simply specifies a mapping from the (simultanous) inputs of the agents to an outcome, often called the *allocation*. Typically, we think of the designer as having some social goal which depends

on some private information held by each agent, called the agent's *type*. For example, in a standard single-item auction, the designer may wish to give the item to the person who values it the most, but cannot do this directly as the desired outcome depends on the agents' private preferences.

To obtain the relevant information for her choice, the designer must leverage *incentives* of the agents. That is, the designer must capitalize on the fact that agents have some stake in the allocation; once chosen, the mechanism's allocation translates to some *utility* of each agent, called the *valuation*, which depends both on the allocation and the agent's private type. In most situations we will consider, part of this allocation is a debt of money, called the *payment* to the mechanism, and it is common to make a *quasi-linear* assumption, that an agents utility of some outcome and payment is the valuation of the outcome minus the payment (see Definition 2.3). We will henceforth refer to the non-payment part of the outcome of the mechanism as the allocation.

What the designer ultimately seeks then is a mechanism with two important properties: (1) *direct*, meaning agents directly input a type to the mechanism, and (2) *incentive compatible* or *truthful*, in that it is in the best interest of each agent to report their true type. While "best interest" is yet undefined, typically truthfulness means that no matter what the other agents report, the remaining agent is no better off lying than reporting his true type. Assuming then that each agent is rational, such a mechanism would provide the designer with all relevant information to make her choice.

Do any truthful mechanisms exist? In fact, they are easy to construct. Suppose for simplicity that there is just a single agent. For each outcome $o \in \mathcal{O}$, choose some arbitrary price $p(o)$. Take each agent's reported type θ as input, and select the outcome which maximizes the agent's total utility $v(\theta, o) - p(o)$, where v is the agent's valuation function. Clearly, as the mechanism is already optimizing on behalf of the agent, the agent can only hurt himself by misreporting his type. Amazingly, the above scheme, suitably generalized to multiple agents, describes *every* truthful direct mechanism (cf. [77]). It goes by many names—the direct characterization, the taxation principle, or the menu auction—and will come up often throughout the dissertation.

But what about the designer? In the above, we just saw how to gain the private information from the agents, but in mechanism design, unlike in scoring rules or prediction markets, typically we are not collecting information for information's sake, but to achieve some goal of the designer which merely *depends* on this information. Thus, a central question in mechanism design is that of *implementability*: given some allocation function f mapping agents' types to a desired outcome, is there a payment function p which when combined with f yields a truthful mechanism?

As it turns out, if the designer is interested in optimizing social welfare, she is in luck: any allocation rule which selects the outcome maximizing the (weighted) sum of agents' valuations can be implemented. This is a classic result due to Vickrey [97], Clarke [37], and Groves [53], called the *VCG mechanism*. The key insight is the form of the payments making the social optimum truthful, which intuitively charge each agent his or her *externality* imposed on the other agents — in other words, each agent i must pay the mechanism what

i costs the other agents by simply participating. (See §1.1.2 for an example.) Surprisingly, Roberts [85] showed that if the mechanism could potentially face agents with arbitrary valuations, then (weighted) VCG mechanisms are the *only* truthful mechanisms.

As the direct characterization given above is often hard to work with when designing mechanisms, much work in the field of mechanism design has been toward finding simpler and more practical implementability conditions. The most famous one is due to Myerson [72], which states that for a "single-parameter domain," where types are real-valued, an allocation function is implementable if and only if it is *monotone* for each agent. For example, when auctioning off bushels of corn, the number of bushels allocated to an agent (fixing the reports of all other agents) must be non-decreasing in the agent's report. In higher dimensions, one must appeal to more complicated versions of monotonicity, like *cyclic monotonicity (CMON)*, which Rochet [86] adapted from convex analysis (cf. Rockafellar [89]). We will explore these generalizations of monotonicity in great detail in Chapter 5.

Once one has decided that an allocation rule is implementable, it of course remains to find payments that render it truthful. Luckily, Myerson came to the rescue here as well, with what is now called the *revenue equivalence* theorem: the payment rule is uniquely determined up to a constant by the allocation rule. Note however that this "constant" may depend on the reports of other agents. We will return to revenue equivalence in §5.3.

Finally, thus far we have considered only direct mechanisms, which raises the question: can the designer gain something by asking agents something other than their raw type? The answer turns out to be no. This is another classic result, called the *revelation principle*, due to Gibbard [49] and later extended by Myerson [72, Lemma 1] and others [74]. The idea is simple: if the input space were different from the type space, than any equilibrium of the mechanism could be replicated by another mechanism which took the agents' actual types and selected the optimal report in the original mechanism. Hence, without loss of generality, one can consider only direct mechanisms. The revelation principle will come up several times in Chapter 3, where we extend our model to accept reports from an alternate space.

1.3 The importance of being characterized

The central focus of this dissertation is that of *characterization*: given a particular setting, what are all truthful elicitation schemes or mechanisms? Of course, this is but one of many questions one could ask, yet we deem this question the most fundamental and fruitful, for reasons we now motivate.

1.3.1 Conceptual

We wish to characterize elicitation models in part because we seek deeper understanding of what it means to be truthful. Can we paraphrase the raw truthfulness constraints in a way that more clearly illuminates what can and cannot be done? Can we cast elicitation in a light that enables new insights and intuition? In fact, we attempt just that by appealing to

the geometric language of convex analysis. We will put this geometric perspective to use in Chapter 3, and to a lesser extent Chapters 4 and 5.

Beyond building intuition and constructing useful representations, we wish to see the *relationships* among the different models. Scoring rules and prediction markets both elicit probabilistic information from agents, yet seem to take wildly different approaches to doing so — are these models related in any way, and if so, how? Are scoring rules more or less expressive than prediction markets? Using characterizations for each model, we will answer these questions and more in Chapter 4. We also explore connections between scoring rules and mechanism design in Chapter 2.

1.3.2 Aligning incentives

Another reason to characterize is to account for what happens off the beaten path of Nash equilibrium. While we often assume that agents will follow their own interests, the precise incentives *leading* them there matter. One reason is that our models are imperfect: we cannot hope to capture *all* incentives confronting an agent, or that agents will be perfectly rational, and in these cases, we may want to make sure that the magnitude of the incentives is strong enough to (mostly) counteract such errors (see e.g. [45]). In this subsection we focus on another situation, when the principal herself is aquiring information merely as a means to an end; she has her own utility function over outcomes of the mechanism or over the information gathered, and wants to *align* the incentives of the elicitation to match her own stake in the game.

Characterizations play a big role when reasoning about all payoffs at once, both in and out of equilibrium. We would not be content with a single elicitation scheme which guarantees that the agent will be truthful in equilibrium; rather, we would like to look at *all possible* schemes, which by definition yield the same equilibrium behavior, and select the one whose entire incentive structure fits best with our goals. For example, suppose it costs an agent \$1 to provide each bit of accuracy; then if $L(r', r)$ is the cost of the report r' to the principal when the correct report is r, the principal may want to encode her own incentives in the payoffs of the mechanism itself, to offset the cost to the agents for the desired accuracy. But for which L can this be done? A characterization is exactly what we need to answer this question.

To illustrate the importance of aligning incentives, we give a third parable, motivated by a well-known problem in information theory.

Case study: compressing a data stream

Imagine a firm is looking to do *compression* on an unfamiliar channel, and from this channel the firm will receive a stream of m characters from an n-sized alphabet which we will index by $[n]$. The goal is to select a binary encoding of this alphabet to minimize the total bits required to store the data, as a cost of \$1 is required for each bit.

A first-order approach to encode such a stream is to assign a probability distribution $p \in \Delta_n$ to the alphabet, and to select an encoding of character i with a binary word of length $\log(1/p(i))$ (we ignore round-off for simplicity). This can be achieved using Huffman Codes for example, and we refer the reader to Cover and Thomas ([38, §5]) for more details. Thus, given a distribution p, the firm pays $L(p;i) = -\log p(i)$ for each character i. It is easy to see that if the characters are sampled from some "true" distribution p^*, then the expected cost $L(p;p^*) := \mathbb{E}_{i\sim p}[L(p;i)] = \mathrm{KL}(p^*;p) + H(p^*)$, which is minimized at $p = p^*$. Not knowing the true distribution p^*, the firm is thus interested in finding a p with a low expected cost $L(p;p^*)$.

An attractive option available to the firm is to *crowdsource* the task of lowering this cost $L(\cdot;\cdot)$ by setting up a prediction market. It is reasonably likely that outside individuals have private information about the behavior of the channel and, in particular, may be able to provide a better estimate p of the true distribution of the characters in the channel. As just discussed, the better the estimate the cheaper the compression.

The firm takes the automated prediction market maker approach, as introduced in §1.2.2. The firm announces that it will select a character from the stream *uniformly at random*, and offers to buy or sell securities, one for each character i of the stream alphabet, which pay out \$1 if i is selected and \$0 otherwise. For the cost function C, the firm chooses the LMSR, namely $C(q) = \log \sum_{i=1}^{n} \exp(q_i)$.

We have devised this payout scheme according to the selection of a single character i, and it is worth noting that because this character is sampled uniformly at random from the stream (with private randomness), the participants *cannot* know which character will be released. This forces the participants to wager on the empirical distribution $\hat{p}$ of the characters from the stream. While this approach has a high variance in the payout, the firm can lower the variance by averaging the payout over all, or a large subset, of the stream. That is, by paying out a $1/k$ fraction for each of k i.i.d. samples of the stream. Given the above the discussion, the firm sets the initial share vector q_0 so that the prices reflect the firm's guess at the stream's empirical distribution, namely some $p_0 \in \Delta_n$.

We may naturally wonder: how does this prediction market benefit the firm that wants to design the encoding? More precisely, if the firm uses the final prices p_T of the market, instead of its initial guess p_0, what is the trade-off between the money paid to participants and the money gained by using the crowdsourced hypothesis? At first glance, it appears that this trade-off can be arbitrarily bad: the worst case cost of encoding the stream using the final estimate p_T is $\sup_{i,p_T} -\log(p_T(i)) = \infty$. Amazingly, however, by virtue of the aligned incentives, the firm has a very strong control of its total cost (the prediction market payout cost plus the encoding cost). Suppose the firm scales the prediction market payouts by a parameter α, to separate the scale of the market from the scale of the encoding cost (which recall is \$1 per bit). Then given any initial estimate p_0 and final estimate p_T, the expected

total cost over p^* is

$$\begin{aligned}\text{Total expected cost} &= \overbrace{m\left(H(p^*) + \mathrm{KL}(p^*; p_T)\right)}^{\text{Encoding cost of using } p_T \text{ given } p^*} + \overbrace{\alpha\left(\mathrm{KL}(p^*; p_0) - \mathrm{KL}(p^*; p_T)\right)}^{\text{Total market payoff for getting advice } p_T} \\ &= m\,H(p^*) + (m-\alpha)\mathrm{KL}(p^*; p_T) + \alpha\,\mathrm{KL}(p^*; p_0)\end{aligned}$$

Let us spend a moment to analyze the above expression. Imagine that the firm set $\alpha = m$, the total number of characters to be encoded. Then the total cost of the firm would be $m\left(H(p^*) + \mathrm{KL}(p^*; p_0)\right)$, which is bounded by $m \log n$ for p_0 uniform. Notice that this expression *does not depend on* p_T – in fact, this cost precisely corresponds to the scenario where the firm had not set up a prediction market and instead used the initial estimate p_0 to encode. In other words, for $\alpha = m$, the firm is *entirely neutral* to the quality of the estimate p_T; even if the market provided an estimate p_T which performed significantly worse than p_0, the cost increase due to the bad choice of p is recouped by payments from the ill-informed participants.

The firm may not want to be neutral to the estimate of the crowd, however, and under the reasonable assumption that the final estimate p_T will improve upon p_0, the firm should set $0 < \alpha < 1$ (of course, positivity is needed for nonzero payouts). In this case, the firm will strictly gain by running the market when $\mathrm{KL}(p^*; p_T) < \mathrm{KL}(p^*; p_0)$, but still has some insurance policy if the estimate p_T is poor.

1.3.3 Impossibility

Characterizations are as much about what is possible as what is impossible. After convincing ourselves that elicitation is possible in certain settings, we naturally begin to wonder what the limitations are. What is the upper limit on what can be done?

Such questions are forefront in the field of algorithmic mechanism design, where computational hardness results have left researchers wondering when a mechanism can approximate some desired outcome in a truthful way. Showing that in some cases one *cannot* achieve such truthful approximations, however, requires a characterization. Roberts' theorem (see §1.2.3) and variants thereof are currently the main tools here, but characterizations in restricted domains would allow us to draw the line of computational feasibility more precisely.

1.4 Overview and organization

We present our model of truthful elicitation in Chapter 2, and show that it generalizes and extends both mechanisms and scoring rules, as well as several other models which have appeared in the literature. Immediately we obtain results in both scoring rules and mechanism design: a characterization of proper scoring rules for non-convex sets of distributions, and a relaxation of an outcome compactness assumption needed by Archer and Kleinberg [7] for allocation rules on non-convex type spaces.

We then turn in Chapter 3 to analyzing situations where, rather than asking for an agent's type, we wish to elicit a simpler representation of it. In the scoring rules context, this has been studied as the elicitation of properties or statistics of a distribution, such as the mean or median [92, 65, 50]. In mechanism design, this is implicit in settings such as matching, where a ranking over potential matches is elicited rather than the agent's utility for them. We extend our model to this setting, and develop a corresponding characterization theorem which generalizes our main result. We find that in essence properties are nothing more than subgradient mappings of convex functions in disguise, and use this insight to develop very general results. For example, we easily see that if a set of types or distributions has positive measure in its convex hull then then there is a unique value of the property almost everywhere. Finally, in §3.3 we further explore notions of duality which stem from convex duality, and in §3.4 provide novel characterizations for the finite-valued, linear, and smooth nonlinear cases.

In Chapter 4 we put our property results to use, showing a strong four-way equivalence between Bregman divergences, scoring rules for linear properties, incomplete prediction markets, and (generalized) exponential families. We show that each pair of these concepts has its own interesting story, and explore these connections to give new insights; for example, an incomplete prediction market can be expressed as a complete prediction market whose prices are constrained to a generalized exponential family of distributions, whose statistic is equal to the payoff function of the incomplete market.

We turn to mechanism design in Chapter 5, where we examine a number of characterizations of when there exist payments that make a given allocation rule truthful. We demonstrate the fundamental role of convexity in mechanism design by rephrasing these results in terms of convex analysis, and instead of appealing to the commonly-used implementability condition known as *cyclic monotonicity*, we directly apply the condition of being a subgradient. This yields substantially simpler and more constructive proofs. We go on to demonstrate how the revenue equivalence theorem from mechanism design falls out almost automatically from our characterization, and apply results for finite-valued properties to produce a novel (non-)implementability check: in mechanisms which select among a finite set of (allocation,payment) pairs, the sets of types that select each outcome are not just polyhedral but form a power diagram.

We conclude in Chapter 6 with a study of the interpretation of prediction market prices. Building on recent connections between prediction markets and learning, we show that the standard automated market makers are in essence performing stochastic mirror descent when trader demands are sequentially chosen at random from a fixed distribution. This provides new insights into how market prices (and price paths) may be interpreted as a summary of the market's belief distribution, by relating them to the optimization problem being solved. In particular, we show that under certain conditions, the stationary point of the stochastic price process generated by the market is equal to the market's Walrasian equilibrium from classic market analysis. Together, these results suggest how traditional market making mechanisms might be replaced with general purpose learning algorithms while still retaining guarantees about their behavior.

1.4.1 Bibliographic remarks

Much of this dissertation is based on work developed in collaboration with coauthors. Chapters 2, 3, and 5 are based on joint work in progress with Ian Kash [47], though large parts of §3.4.2 and §3.4.3 were discovered with Jacob Abernethy. Chapter 4 is drawn from work with Jacob Abernethy [4, 2] and ongoing work with Mark Reid and Robert Williamson. Finally, Chapter 6 is adapted from joint work with Mark Reid and Nicolás Della Penna [46].

Chapter 2

Unifying elicitation via convex analysis

In Chapter 1, we discussed several elicitation models and settings, where an agent has private information and a mechanism wishes to extract the information via incentives. In this chapter and the next, we introduce a new model of elicitaton which generalizes all of these settings using tools from convex analysis. In our model, a single agent is endowed with some type t that is private information and is asked to reveal it. After doing so, she receives a score that depends on both her report t' and her true type t.

For reasons that will become clear, we represent this as a function $\mathfrak{A}(t')(t)$ that maps her reported type to a function that maps types to real numbers, with her score being this function applied to her true type (equivalently her reported type selects from a parameterized family of functions with the result applied to her true type). We allow $\mathfrak{A}$ to be quite general, with the main requirement being that $\mathfrak{A}(t')(\cdot)$ is an affine[1] function of the true type t, and seek to understand when it is optimal for the agent to truthfully report her type. Given this restriction, it is immediately clear why convexity plays a central role — when an agent's type is t, the score for telling the truth is $\mathfrak{A}(t)(t) = \sup_{t'} \mathfrak{A}(t')(t)$, which is a convex function of t as the pointwise supremum of affine functions.

We first discuss the role of convexity in mechanisms, scoring rules, and prediction markets. We then state our model and result, and show how many existing elicitation models are special cases of our model.

2.1 Convexity in elicitation

Viewing our model as a mechanism, $\mathfrak{A}(t')$ can be thought of as the allocation and payment given a report of t', which combine to determine the utility of the agent as a function of her

[1] A mapping between two vector spaces is affine if it consists of a linear transformation followed by a translation.

type.[2] In this context, $\mathfrak{A}(t)(t)$ is the consumer surplus function, and Myerson's well-known characterization [72] states that, in single-parameter settings, a mechanism is truthful if and only if the consumer surplus function is convex and its derivative (or subgradient at points where it is not differentiable) is the allocation rule; we formalize this as Theorem 2.3. More generally, this remains true in higher dimensions (cf. [7]). Note that here the restriction that $\mathfrak{A}(t')$ be affine is without loss of generality, since we may consider types as functions mapping an outcome to the agent's utility for that outcome, and the evaluation of a type on an outcome (or a distribution over outcomes) is affine in that type.

We may also view $\mathfrak{A}$ as a scoring rule, where an agent is asked to predict the distribution of a random variable and given a score based on the observed realization of that variable. In this setting, types are distributions over outcomes, and $\mathfrak{A}(t')(t)$ is the agent's expected score for a report that the distribution is t' when she believes the distribution is t. As an expectation, this score is linear in the agent's type, and hence affine.[3]

The role of convexity in the scoring rules literature dates back to 1956, when John McCarthy [71] observed that scoring rules could be derived as a generalized derivative of a convex function. This was clarified and extended by Savage in 1971, who showed explicitly how to construct a scoring rule from a convex function (the negative of the *Bayes risk* from decision theory) [92]. The general, modern characterization of scoring rules is due to Gneiting and Raftery [51], who used a more nuanced convex analysis approach to clarify and generalize a number of previous characterizations, including Savage [92] and Schervish [93].

We will see in Chapter 3 and Chapter 4 how prediction markets may be encorporated into our model, but for now we merely underscore the role of convexity. As we saw in §1.2.2, the now-standard automated prediction market model of Abernethy et al. [1] uses a convex cost potential function C. In essence, this convexity enforces the same kind of monotonicity as in mechanisms — the more a trader buys a particular contract, the higher the price becomes. The convexity of C arises from an axiomatic approach, and is fundamental to the model.

The similarity between scoring rules and mechanism design was noted by Fiat et al. [45], who gave a construction to convert mechanisms into scoring rules and vice versa. In this chapter, we prove a general characterization, of which these characterizations of scoring rules and mechanisms are special cases. Our proof is essentially a combination of Gneiting and Raftery's scoring rule construction [51] with a technique from Archer and Kleinberg [7] for handling mechanisms with non-convex type spaces. Our characterization not only shows how mechanisms and scoring rules relate to each other, but also provides an understanding of how results about mechanisms relate to results about scoring rules and vice versa. In particular, many results in each literature are fundamentally results in convex analysis, and by phrasing them as such it is immediately clear how they apply in the other domain.

[2] It suffices to consider a single agent because notions of truthfulness such as dominant strategies and Bayes-Nash are phrased in terms of holding the behavior of other agents constant. See [35, 7] and §2.3.1 for additional discussion.

[3] Since distributions lie on an affine space, any affine function can be implemented as well.

2.2 Model and main result

To begin, we introduce the terminology which we use throughout the dissertation; see Table 2.1 for notation. A function $G : X \to \overline{\mathbb{R}}$ is *convex* if $G(\alpha x + (1-\alpha)x') \leq \alpha G(x) + (1-\alpha)G(x')$ for all $x, x' \in X$ and $\alpha \in [0,1]$. The *domain* of a convex G is the set $\mathsf{dom}(G) \doteq \{x \in X : G(x) < \infty\}$. G is a *proper* convex function if it never takes on $-\infty$ and its domain is nonempty. We write $\mathsf{conv}(S)$ to denote the *convex hull* of vector space X, the set of all (finite) convex combinations of elements of S. The *relative interior* of a convex set S is the interior when restricted to the smallest affine subspace containing S; formally, $\mathsf{relint}(S) := \{x \in S : \forall y \in S\ \exists \lambda < 1 : \lambda x + (1-\lambda)y \in S\}$ [103, pp. 2-3]. A topological vector space is *locally convex* if every neighborhood of 0 contains a convex neighborhood of 0 [5, Def. 5.71].

$\overline{\mathbb{R}}$	extended real numbers $\mathbb{R} \cup \{-\infty, \infty\}$
$[n]$	set $\{1, \ldots, n\}$
$\Delta(X)$	set of all probability measures on X
Δ_n	set of all probability measures on $X = [n]$
$\mathbb{1}$	all-ones vector in $\mathbb{R}^X$ with $\mathbb{1}(x) = 1$ for all $x \in X$
$\mathbb{1}_x$	indicator or standard vector in $\mathbb{R}^X$ with $\mathbb{1}_x(x) = 1$ and 0 otherwise
∂	subgradient operator
id_X	identity function on X, $\mathrm{id}_X : x \mapsto x$
$\mathsf{Lin}(X \to Y)$	set of linear functions from X to Y
$\mathsf{Aff}(X \to Y)$	set of affine functions from X to Y
$\mathsf{conv}(S)$	convex hull of S
$\mathsf{relint}(S)$	relative interior of S
$\mathsf{dom}(G)$	domain $\{x \in X : G(x) < \infty\}$ of a convex function G
Eval_o	evaluation operator $\mathsf{Eval}_o[f] = f(o)$
$\mathrm{KL}(p\|q)$	relative entropy $\int_X \log\left(\frac{dp}{dq}\right) dp$

Table 2.1: Notation

Let $\mathcal{T} \subseteq \mathcal{V}$ for some vector space $\mathcal{V}$. We consider a very general model with an agent who has a given type $t \in \mathcal{T}$ and reports some possibly distinct type $t' \in \mathcal{T}$, at which point the agent is rewarded according to some score $\mathfrak{A}$ which is affine in the true type t. This reward we call an affine score. We wish to characterize all *truthful* affine scores, those which incentivize the agent to report her true type t.

Definition 2.1. *Any function $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$, where $\mathcal{T} \subseteq \mathcal{V}$ for some vector space $\mathcal{V}$ and $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{V} \to \overline{\mathbb{R}})$, is an* affine score. *We say $\mathfrak{A}$ is* truthful *if for all $t, t' \in \mathcal{T}$,*

$$\mathfrak{A}(t')(t) \leq \mathfrak{A}(t)(t). \tag{2.1}$$

If this inequality is strict for all t, t', then $\mathfrak{A}$ is strictly truthful.

Our characterization relies heavily on convex analysis, a central concept of which is the subgradient of a function.

Definition 2.2. *Given some function $G : \mathcal{T} \to \mathbb{R}$, a function $dG_t \in \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}})$ is a* subgradient *to G at t if for all $t' \in \mathcal{T}$,*

$$G(t') \geq G(t) + dG_t(t' - t). \tag{2.2}$$

We denote by ∂G_t or $\partial G(t)$ the set of subgradients to G at t, and $\partial G = \cup_{t \in \mathcal{T}} \partial G_t$.

Before stating our characterization, we first must address the role of $\pm\infty$ as a viable payoff. For mechanism design, it is typical to assume that utilities are always real-valued. However, the log scoring rule (one of the most popular scoring rules) has the property that if an agent reports that an event has probability 0, and then that event does occur, the agent receives a score of $-\infty$. Essentially solely to accommodate this, we allow affine scores and subgradients to take on values from the extended reals.

Once one allows algebra on the extended reals, great care must be taken to avoid the indeterminate expression $\infty - \infty$. To address this point, it is standard (cf. [51]) to restrict consideration to the "regular" case, where intuitively only things like the log score are permitted to be infinite. Formally, we say a parameterized family of linear functions (e.g. a family of subgradients) $\{dG_t \in \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}})\}_{t \in \mathcal{T}}$ is *$\mathcal{T}$-regular* if $dG_t(t) \in \mathbb{R}$ for all $t \in \mathcal{T}$, and $dG_{t'}(t) \in \mathbb{R} \cup \{-\infty\}$ for $t' \neq t$.[4] Likewise, $\mathcal{T}$-regular affine functions $\{a_t \in \mathsf{Aff}(\mathcal{V} \to \overline{\mathbb{R}})\}_{t \in \mathcal{T}}$ have $a_t(t) \in \mathbb{R}$ for all $t \in \mathcal{T}$, and $a_t(t') < \infty$ for $t' \neq t$. In particular, an affine score $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$ is *regular* if $\mathfrak{A}(t)(t) \in \mathbb{R}$ for all $t \in \mathcal{T}$, and $\mathfrak{A}(t')(t) \in \mathbb{R} \cup \{-\infty\}$ for $t' \neq t$.

For the remainder of the dissertation we assume all affine scores and parameterized families of linear or affine functions are $\mathcal{T}$-regular, where $\mathcal{T}$ will be clear from context. Note that certain results in the following three chapters require a stronger assumption that the relevant parameterized families are in fact real-valued rather than simply regular. A reader not interested in the details of how our framework incorporates the log scoring rule can assume that all affine scores and families are real-valued throughout the dissertation with little loss.

We now state, and prove, our characterization theorem. The proof draws techniques and insights from Gneiting and Raftery [51] and Archer and Kleinberg [7], but in such a way as to simplify the argument considerably.

Theorem 2.1. *Let regular affine score $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$ be given. $\mathfrak{A}$ is truthful if and only if there exists some convex $G : \mathsf{conv}(\mathcal{T}) \to \overline{\mathbb{R}}$ with $G(\mathcal{T}) \subseteq \mathbb{R}$, and some selection of subgradients $\{dG_t\}_{t \in \mathcal{T}}$, such that*

$$\mathfrak{A}(t')(t) = G(t') + dG_{t'}(t - t'). \tag{2.3}$$

[4] To define linear functions to $\overline{\mathbb{R}}$, we adopt the convention $0 \cdot \infty = 0 \cdot (-\infty) = 0$. Thus, any $\ell \in \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}})$ can be written as $\ell_1 + \infty \cdot \ell_2$ for some $\ell_1, \ell_2 \in \mathsf{Lin}(\mathcal{V} \to \mathbb{R})$. Similarly, $\mathsf{Aff}(\mathcal{V} \to \overline{\mathbb{R}}) = \{t \mapsto \ell(t - v) + c \mid \ell \in \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}}), v \in \mathcal{V}, c \in \mathbb{R}\}$, representing possibly vertical hyperplanes.

Proof. It is trivial from the subgradient inequality (2.2) that the proposed form is in fact truthful. For the converse, we are given some truthful $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$. Note first that for any $\hat{t} \in \mathsf{conv}(\mathcal{T})$ we may write $\hat{t}$ as a finite convex combination $\hat{t} = \sum_{i=1}^{m} \alpha_i t_i$ where $t_i \in \mathcal{T}$. Now, as the range of $\mathfrak{A}$ is affine, we may naturally extend $\mathfrak{A}(t)$ to all of $\mathsf{conv}(\mathcal{T})$ by defining

$$\mathfrak{A}(t)(\hat{t}) = \sum_{i=1}^{m} \alpha_i \mathfrak{A}(t)(t_i). \tag{2.4}$$

One easily checks that this definition coincides with the given $\mathfrak{A}$ on $\mathcal{T}$.

Now we let $G(\hat{t}) \doteq \sup_{t \in \mathcal{T}} \mathfrak{A}(t)(\hat{t})$, which is convex as the pointwise supremum of convex (in our case affine) functions. Since $\mathfrak{A}$ is truthful, we in particular have $G(t) = \mathfrak{A}(t)(t) \in \mathbb{R}$ for all $t \in \mathcal{T}$ by our regularity assumption. Also by truthfulness, we have for all $t' \in \mathcal{T}$ and $\hat{t} \in \mathsf{conv}(\mathcal{T})$,

$$\begin{aligned} G(\hat{t}) = \sup_{t \in \mathcal{T}} \sum_{i=1}^{m} \alpha_i \mathfrak{A}(t)(t_i) \geq \sum_{i=1}^{m} \alpha_i \mathfrak{A}(t')(t_i) &= \mathfrak{A}(t')(t') + \sum_{i=1}^{m} \alpha_i \mathfrak{A}_\ell(t')(t_i - t') \\ &= G(t') + \mathfrak{A}_\ell(t')(\hat{t} - t'). \end{aligned}$$

Hence, $\mathfrak{A}_\ell(t')$ satisfies (2.2) for G at t', so $\mathfrak{A}$ is of the form (2.3). □

2.3 Existing models as special cases

We conclude this chapter with several models and applications which demonstrate the power of our framework. In particular, as we will see, both scoring rules and mechanisms fit comfortably within our framework. We defer treatment of prediction markets to Chapter 4, as we will need the notion of indirect elicitation which we introduce in Chapter 3.

2.3.1 Mechanism design

We will now show how to view a mechanism, introduced in §1.2.3, as an affine score. First, we formally introduce mechanisms, in the single agent case (see below for remarks about multiple agents).

Definition 2.3. *Given outcome space $\mathcal{O}$ and a type space $\mathcal{T} \subseteq (\mathcal{O} \to \mathbb{R})$, consisting of functions mapping outcomes to reals, a (direct)* mechanism *is a pair (f, p) where $f : \mathcal{T} \to \mathcal{O}$ is an* allocation rule *and $p : \mathcal{T} \to \mathbb{R}$ is a* payment*. The utility of the agent with type t and report t' to the mechanism is $U(t', t) = t(f(t')) - p(t')$; we say the mechanism (f, p) is* truthful *if $U(t', t) \leq U(t, t)$ for all $t, t' \in \mathcal{T}$.*

Here we suppose that the mechanism can choose an allocation from some set $\mathcal{O}$ of outcomes, and there is a single agent whose type $t \in \mathcal{T}$ is itself the valuation function. That is, the agent's net utility upon allocation o and payment p is $t(o) - p$. We may view the

type space $\mathcal{T}$ as lying in the vector space $\mathcal{V} = \mathbb{R}^{\mathcal{O}}$: by definition, for any $v_1, v_2 \in \mathcal{V}$, we have $(v_1 + \alpha v_2)(o) = v_1(o) + \alpha v_2(o)$, so $v_1 + \alpha v_2 \in \mathcal{V}$. So while we have made no assumptions about $\mathcal{O}$ or the form of v_1 and v_2, this function application, called the *evaluation operator* $\mathsf{Eval}_o[v] \doteq v(o)$, is a linear operation. Thus, given any outcome o and constant c, the mapping $t \mapsto t(o) + c$ is an affine function from $\mathcal{T}$ to $\mathbb{R}$. In particular, this holds for the utility of an agent $U(t', \cdot) = \mathsf{Eval}_o[t] - p(o)$ given that our mechanism chooses outcome $f(t') = o$. (While according to our definition p depends directly on the report t', it is easy to see that for any truthful mechanism, p depends on t' only through the allocation $o = f(t')$; otherwise agents with $t \in f^{-1}(o)$ would always report $\operatorname{argmin}_{t: f(t)=o} p(t)$.) Thus, we have an affine score $\mathfrak{A}(t')(t) \doteq U(t', t)$, where $\mathcal{A} = \{t \mapsto t(o) + c \,|\, o \in \mathcal{O}, c \in \mathbb{R}\}$, so that every combination of outcome and payment a mechanism could choose is an element of $\mathcal{A}$.

Merely as an illustration, we will use Theorem 2.1 to show two simple characterizations that have appeared in the mechanism design literature. The first is the *direct characterization*, which simply states that a mechanism must optimize on behalf of the agent. More precisely, we consider the mechanism which assigns a price $p(o) \in \mathbb{R}$ to each outcome and chooses the outcome and corresponding price that maximizes the agent's welfare:

$$f(t) \in \operatorname{argsup} \{t(o) - p(o) : o \in \mathcal{O}\}. \tag{2.5}$$

We show that this mechanism is truthful, and moreover every truthful mechanism can be represented this way.

For the forward direction, we pick prices $p : \mathcal{O} \to \mathbb{R}$ and set $G(t) \doteq \max_{o \in \mathcal{O}} t(o) - p(o)$.[5] As G is a pointwise maximum of affine functions, it is convex. For a convex function defined in this manner, it is easy to verify that if $o^* \in \operatorname{argmax}_{o \in \mathcal{O}} t^*(o) - p(o)$, then Eval_{o^*} is a subbgradient of G at t^*:

$$\begin{aligned} G(t^*) + \mathsf{Eval}_{o^*}[t - t^*] &= \max_{o \in \mathcal{O}} \{t^*(o) - p(o)\} + t(o^*) - t^*(o^*) \\ &= t(o^*) - p(o^*) \leq \max_{o \in \mathcal{O}} t(o) - p(o) = G(t). \end{aligned}$$

Thus, letting $f(t) = \operatorname{argmax}_{o \in \mathcal{O}} t(o) - p(o)$ for all t, we can apply Theorem 2.1 to show that the affine score $\mathfrak{A}(t')(t) = t(f(t')) - p(f(t'))$ is truthful.

For the converse, let truthful mechanism f, p be given. As argued above, p must depend on t' only through $f(t')$, and hence we may re-represent it as a function $p : \mathcal{O} \to \mathbb{R}$ so that $U(t', t) = t(f(t')) - p(f(t'))$. But then by the proof of Theorem 2.1 we have $U(t, t) = \max_{t'} t(f(t')) - p(f(t')) = \max_o t(o) - p(o)$ as desired.

The above discussion could have been greatly simplified by appealing to the following simple fact about affine scores:

Proposition 2.2. *An affine score $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$ is truthful if and only if*

$$\mathfrak{A}(t) \in \operatorname{argsup} \{a(t) : a \in \mathfrak{A}(\mathcal{T})\}. \tag{2.6}$$

[5] Strictly speaking we need to appropriately restrict $\mathcal{T}$ or $\mathcal{O}$ so that this in fact a maximum rather than a supremum for our VCG-style mechanism of giving the agent her preferred outcome to even be well defined.

The proof is trivial and follows directly from Definition 2.1. Unfolding the result and using the definition of a mechanism then yields the form (2.5), again appealing to the fact that only the lowest price for a given outcome will ever be chosen.

The result that all mechanisms have the form (2.6) is sometimes called the *taxation principle*, and the form of the mechanism is called the *menu auction*, referring to the "menu" of allocations and prices $\{(o, p(o))\}$ available to the agent. This is an important representation for any affine score, as we will see in Chapter 3 and specifically §3.3.2. Interestingly, as we will see, the menu always satisfies $p(o) = G^*(\mathsf{Eval}_o)$, where $G(t) = U(t,t)$ and G^* is the *convex conjugate* of G.

We proceed to our second second characterization, due to Myerson [72], for a single parameter setting (i.e. when the agent's type can be described by a single real number). The result states that an allocation rule is implementable, meaning there is some payment rule making it truthful, if and only if it is *monotone* in the agent's type. We will explore monotonicity conditions in higher dimensions in Chapter 5.

Theorem 2.3 (Myerson [72])**.** *Let $\mathcal{T} = \mathbb{R}_+$, $\mathcal{O} \subseteq \mathbb{R}$, so that the agent's valuation is $t \cdot o$. Then a mechanism f, p is truthful if and only if*

i. f is monotone non-decreasing in t,

ii. $p(t) = tf(t) - \int_0^t f(t')dt' + p_0$.

Proof. By elementary results in convex analysis f is a subgradient of a convex function on $\mathbb{R}$ if and only if it is monotone non-decreasing. By Theorem 2.1, the mechanism is truthful if and only if f is the subgradient of the particular function $G(t) = U(t,t) = tf(t) - p(t)$, which is equivalent to (i) and the condition $G(t) = \int_0^t f(t')dt' + C$. □

Finally, we remark on what may appear as limitations in our approach. First, note that we have focused on the *single-agent* case here, even though much of the mechanism design literature addresses the multi-agent case. In some sense, extending our characterizations to multiple agents is trivial: a mechanism is truthful if and only if it is truthful for agent i when fixing the reports of the other agents. Hence, we merely apply our characterization to each single-agent mechanism induced by reports of the other agents. This is sufficient for our present study, but there are certainly reasons to take a more nuanced approach to the multi-agent setting — see §5.4 for further discussion.

Another apparent limitation is that we are locked into a deterministic and non-Bayesian setting. This is purely for ease of exposition; if one is interested in randomized mechanisms, one can take $f : \mathcal{T} \to \Delta(\mathcal{O})$ and define $U(t', t) = \mathbb{E}_{o \sim f(t')}[t(o)] - p(t')$, which is still affine in t. Even if one does not assume risk-neutral agents, taking the outcome space to be $\mathcal{O}' \doteq \Delta(\mathcal{O})$ is sufficiently general. Finally, Bayesian agents can also be represented; in the above discussion of the multi-agent setting, take expectations instead of fixing specific values for the other agents.

2.3.2 Scoring rules for non-convex $\mathcal{P}$

In this section, we show that the Gneiting and Raftery characterization is a simple special case of Theorem 2.1, and moreover that we *generalize* their result to the case where the set of distributions $\mathcal{P}$ may be non-convex. We also mention a result about local properness using tools we will develop in §5.1.2. To begin, we formally introduce scoring rules and show that they fit into our framework.

Definition 2.4. *Given outcome space $\mathcal{O}$ and set of probability measures $\mathcal{P} \subseteq \Delta(\mathcal{O})$, a* scoring rule *is a function $\mathfrak{S} : \mathcal{P} \times \mathcal{O} \to \overline{\mathbb{R}}$. We say $\mathfrak{S}$ is* proper *if for all $p, q \in \mathcal{P}$,*

$$\mathbb{E}_{o\sim p}[\mathfrak{S}(q, o)] \leq \mathbb{E}_{o\sim p}[\mathfrak{S}(p, o)]. \tag{2.7}$$

If the inequality in (2.7) is strict, then $\mathfrak{S}$ is strictly proper*.*

Just as above for mechanisms, the type space here is trivial: $\mathcal{T} = \mathcal{P}$. Thus, we need only construct the correct set of affine functions available to the scoring rule as payoff functions. In this case, letting $\mathcal{F}$ be the set of $\mathcal{P}$-quasi-integrable[6] functions $f : \mathcal{O} \to \overline{\mathbb{R}}$, we simply choose $\mathcal{T} = \mathcal{P}$ and $\mathcal{A} = \{p \mapsto \int_{\mathcal{O}} f(o)\, dp(o) \mid f \in \mathcal{F}\}$. Note that in this case $\mathcal{A}$ actually contains *linear* functions of p, which are trivially affine. Thus, a scoring rule $\mathfrak{S}$ is an affine score $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$ for the above choices of $\mathcal{T}$ and $\mathcal{A}$.

We now apply Theorem 2.1 for our choice of $\mathcal{T}$ and $\mathcal{A}$, which yields the following generalization of Gneiting and Raftery [51].

Corollary 2.4. *For an arbitrary set $\mathcal{P} \subseteq \Delta(\mathcal{O})$ of probability measures, a regular[7] scoring rule $\mathfrak{S} : \mathcal{P} \times \mathcal{O} \to \overline{\mathbb{R}}$ is proper if and only if there exists a convex function $G : \mathsf{conv}(\mathcal{P}) \to \mathbb{R}$ with functions $G_p \in \mathcal{F}$ such that*

$$\mathfrak{S}(p, o) = G(p) + G_p(o) - \int_{\mathcal{O}} G_p(o)\, dp(o), \tag{2.8}$$

where $G_p : q \mapsto \int_{\mathcal{O}} G_p(o)\, dq(o)$ is a subgradient of G for all $p \in \mathcal{P}$.

Proof. The given form is truthful by the subgradient inequality. Let $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$ be a given truthful affine score. Since $\mathfrak{A}(p) \in \mathcal{A}$, we have some $f_p \in \mathcal{F}$ generating $\mathfrak{A}(p)$. We can therefore use $G_p : q \mapsto \int_{\mathcal{O}} f_p(o)\, dq(o)$ as the subgradients in the proof of Theorem 2.1, thus giving us the desired form. □

Importantly, Corollary 2.4 immediately generalizes [51] to the case where $\mathcal{P}$ is not convex, which is a new result to the scoring rules literature. Previously, in the absence of a characterization, several authors have nonetheless worked in the non-convex $\mathcal{P}$ case. For example, Babaioff et al. [13] examine when proper scoring rules can have the additional property that

[6] We say that $f : \mathcal{O} \to \overline{\mathbb{R}}$ is $\mathcal{P}$-quasi-integrable if $\int_{\mathcal{O}} f(o) dp(o) \in \overline{\mathbb{R}}$ for all $p \in \mathcal{P}$.

[7] This is the same concept as with affine scores: scores cannot be ∞ and only incorrect reports can yield $-\infty$.

uninformed experts do not wish to make a report (have a negative expected utility), while informed experts do wish to make one. They show that this is possible in some settings where the space of reports is not convex. Our characterization shows that, despite not needing to ensure properness on on reports outside $\mathcal{P}$, essentially the only possible scoring rules are still those that are proper on all of $\Delta(\mathcal{O})$. Similarly, Fang et al. [43] find conditions on $\mathcal{P}$ for which every continuous "value function" $G : p \mapsto \mathfrak{S}(p, p)$ on $\mathcal{P}$ can be attained by some $\mathfrak{S}$. Given that a convex function can take on *arbitrary* values on the boundary of a strictly convex set (e.g. if it takes on $-\infty$ on the relative interior), it would seem that our approach would provide insights to that question as well.

We will explore local truthfulness conditions in §5.1.2, where one verifies that an affine score is truthful by checking that it is truthful in a small neighborhood around every point. While this is a natural property to examine in mechanism design, it is a less common concern for scoring rules designers, as they tend to operate under fewer constraints than mechanism designers. Nevertheless, our results from §5.1.2 will apply, and in particular Corollary 5.4 shows that local properness is equivalent to global properness for scoring rules on convex $\mathcal{P}$.

Corollary 2.5. *For a convex set $\mathcal{P} \subseteq \Delta(\mathcal{O})$ of probability measures, a scoring rule $\mathfrak{S} : \mathcal{P} \times \mathcal{O} \to \overline{\mathbb{R}}$ is proper if and only if it is (weakly) locally proper.*

2.3.3 Decision Rules

Theorem 2.1 also generalizes Gneiting and Raftery's [51] characterization to settings beyond eliciting a single distribution. For example, a line of work has considered a setting where a decision maker needs to select from a finite set $\mathcal{D}$ of decisions and so desires to elicit the distribution over outcomes conditional on selecting each alternative [80, 32, 31]. Since only one decision will be made and so only one conditional distribution can be sampled, simply applying a standard proper scoring rule generally does not result in truthful behavior. Applying Theorem 2.1 to this setting characterizes what expected scores must be, from which many of the results in these papers follow. As it is not our main focus, we refrain from introducing the model necessary to explicitly state a characterization result similar to Corollary 2.4.

2.3.4 Proper losses for partial labels

Several variants of proper losses have appeared in the machine learning literature, one of which is the problem of estimating the probability distribution of labels for an item when the training data may contain several noisy labels, possibly not even including the correct label. (This is frequently the case, for example, when using crowdsourced labels for items.) More formally, one wishes to estimate $p \in \Delta_n$ where the true label $y \in [n]$ is drawn from p. However, instead of observing a sample $y \sim p$ and designing a proper loss $\ell(\hat{p}, y)$, one instead only observes some noisy set of labels $S \subseteq [n]$. Hence, the task is to design a loss $\ell(\hat{p}, S)$ which when minimized over one's data yields accurate estimates of the true p.

Recently this problem was studied in [36] under the assumption that $S \sim q$ where $q = Mp$ for some known $M \in \mathbb{R}^{2^n \times n}$, meaning if the observed label is drawn from p, the noisy set of labels is drawn from Mp (using some indexing of the sets, say lexographical). Cid-Sueiro provides a characterization of all proper losses for this setting, and we merely note that the (negative) payoff $\mathbb{E}_{S\sim Mp}[\ell(\hat{p}, S)] = \ell(\hat{p}, \cdot)^\top Mp$ is linear in the underlying distribution p, so our Theorem 2.1 applies and allows us to recover her first characterization result. Again, we refrain from introducing the model necessary to explicitly state this result. Note that this is essentially a latent observation setting, and the fact that what we observe is a set of labels is in no way necessary — any observed outcome whose distribution has a linear (or affine) relationship with the latent outcome would suffice to apply our theorem.

2.3.5 The Ω branching model

Several mechanism design settings considered in the literature have some form of *exogenous* randomization, in that "Nature" chooses some outcome ω according to some (often unknown) distribution, which may in turn depend on the allocation chosen by the mechanism. Examples include sponsored search auctions [44], multi-armed bandit mechanisms [14], and recent work on daily deals [27]. The work of Cai et al. [27] introduces a very general model for such settings, which we now describe.

Let $\mathcal{O}$ be a set of allocations, and for each allocation o and each agent i, let $\Omega^{i,o}$ be some set of outcomes. Agents each have a valuation function $v^i : \mathcal{O} \to \mathbb{R}$ and a set of beliefs $p^{i,o} \in \Delta(\Omega^{i,o})$ for each allocation $o \in \mathcal{O}$. The mechanism aggregates all of this information into a single allocation o, and additionally choses some payoff function $s^i : \Omega^{i,o} \to \mathbb{R}$, so that the final utility of agent i is $v^i(o) + \mathbb{E}_{p^{i,o}}[s^i]$. A mechanism is truthful if for all values of v and p for the other agents, agent i maximizes her total utility by reporting v^i and $p^i \doteq (p^{i,o})_{o\in\mathcal{O}}$ truthfully. For example, the standard sponsored search setting has $\Omega^{i,o} = \{\text{click}, \text{no click}\}$ for o such that i is allocated a slot, and the probabilities $p^{i,o}$ are assumed to be public knowledge.

We first observe that this model can easily be cast as an affine score, as follows. For simplicity, we fix some agent i and focus on the single-agent case; as discussed several times above, this is essentially without loss of generality. The type space is simply the combined private information of the agent,

$$\mathcal{T} = \left\{ (v, p) : v \in \mathcal{O} \to \mathbb{R},\ p \in \prod_{o\in\mathcal{O}} \Delta(\Omega^{i,o}) \right\}. \tag{2.9}$$

The utility of the agent upon allocation and payoff o, s is simply $v(o) + \mathbb{E}_{p^o}[s] = \mathsf{Eval}_o[v] + s\mathbb{1}_o^\top p$, which is linear in the type $t = (v, p)$ and therefore affine. (Here we represent p as a matrix in $\mathbb{R}^{\mathcal{O}\times\Omega^{i,o}}$ and $s \in \mathbb{R}^{\Omega^{i,o}}$, and define $\mathbb{1}_o$ to be the standard vector with 1 at entry o and 0 elsewhere.) Thus, letting $t = (v, p)$, we can represent this as an affine score:

$$\mathfrak{A}(t')(t) = v(o(t')) + \mathbb{E}_{p^{o(t')}}[s(t')]. \tag{2.10}$$

Motivated by incorporating the utilities of the end consumers in a daily deal setting, Cai et al. [27] ask when one can implement an allocation rule of the form $f(v,p) = \operatorname{argmax}_{o\in\mathcal{O}} v(o) + g^o(p^o)$; in other words, when does there exist some choice of score $s(v,p) \in \mathbb{R}^{\Omega^{i,f(v,p)}}$ making f truthful. They conclude that this can be done if and only if g^o is convex for each $o \in \mathcal{O}$. It is interesting, and perhaps illuminating, to view this question in terms of our affine score framework.

Stepping back for a moment, consider a type space $\mathcal{T} \subseteq \mathcal{V} = \mathcal{V}^X \times \mathcal{V}^Y$ which partitions into two (subsets of) subspaces. We wish to know when a function $f : \mathcal{T} \to \mathsf{Lin}(\mathcal{V}^X \to \mathbb{R})$ is implementable, in the sense that there exists some truthful affine score $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$, $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$, and some $h : \mathcal{T} \to \mathsf{Aff}(\mathcal{V}^Y \to \mathbb{R})$ such that $\mathfrak{A}(t')(t) = f(t')(t^X) + h(t')(t^Y)$, where of course $t = (t^X, t^Y)$. That is, when can we complete the partial "allocation" f into a truthful affine score?

For convenience, for each $a \in \mathcal{A}$ we write $X(a) \in \mathsf{Lin}(\mathcal{V}^X \to \mathbb{R})$ to be the linear part of a on $\mathcal{V}^X$, and $Y(a)$ to be the *affine* part of a on $\mathcal{V}^Y$. Then we have that f is implementable if and only if

$$f(t) \in \operatorname*{argsup}_{x\in X(\mathcal{A})} \left\{ x(t^X) + \sup_{\substack{a\in\mathcal{A} \\ X(a)=x}} \left\{ Y(a)(t^Y) \right\} \right\} \tag{2.11}$$

To see this, for one direction we simply unfold the direct characterization for affine scores, Proposition 2.2, by taking the supremum first over $X(\mathcal{A})$ and then over the rest. For the other direction, note that taking $\mathfrak{A}(t')(t) = f(t')(t^X) + y(t')(t^Y)$ where y is in the argsup of the supremum of eq. (2.11) gives a truthful affine score.

Returning to the setting at hand, let us denote by $a_{o,s} \in \mathcal{A}$ the function $(v,p) \mapsto v(o') + \mathbb{E}_{p^{o'}}[s]$. We now see that $f(v,p)$ is implementable if and only if it satisfies

$$f(v,p) \in \operatorname*{argsup}_{o\in\mathcal{O}} \left\{ v(o) + \sup_{s:a_{o,s}\in\mathcal{A}} \left\{ \mathbb{E}_{p^o}[s] \right\} \right\}. \tag{2.12}$$

Thus, letting $g^o(p^o) = \sup\{\mathbb{E}_{p^o}[s] : a_{o,s} \in \mathcal{A}\}$, we see that g^o is convex as the supremum of affine functions. Moreover, given any collection of convex functions $\{g^o\}_{o\in\mathcal{O}}$, where $g^o : \Delta(\Omega^{i,o}) \to \mathbb{R}$, we can define $S^o \doteq \{\omega \mapsto g(p) + dg(\mathbb{1}_\omega - p) : p \in \mathsf{dom}(g)\}$ and $\mathcal{A} \doteq \{a_{o,s} : o \in \mathcal{O},\, s \in S^o\}$, thus recovering each g^o in the above expression. It then only remains to show that no other nonconvex function can serve in the argsup; for this one may appeal to the argument of Cai et al. [27] which observes that the indifference points between different allocations is fixed, thus determining the function in the argsup up to a constant.

2.4 Discussion

We have presented a model of truthful elicitation which generalizes and extends both mechanisms and scoring rules. On the mechanism design side, we will see in Chapter 5 how our framework provides simpler and more constructive proofs of a number of known results, some

of which (as we saw in Section 2.3.2) lead to new results about scoring rules. We generalize our model in Chapter 3 to allow reports of a succinct representation of an agent's private information rather than that information itself, a topic that has been studied in the scoring rules literature about eliciting properties of distributions.

Our analysis makes use of the fact that $\mathfrak{A}(t')(t)$ is affine in t to ensure that $G(t) = \sup_{t'} \mathfrak{A}(t')(t)$ is a convex function. However, this property continues to hold if $\mathfrak{A}(t')(t)$ is instead a convex function of t. Thus, a natural direction for future work is to investigate characterizations of convex scores. While mechanisms can always be represented as affine functions by taking the types to be functions from allocations to $\mathbb{R}$, it may be more natural to treat the type as a parameter of a (convex) utility function. While many such utility functions are affine (e.g. dot-product valuations), others such as Cobb-Douglas functions are not. Berger, Müller, and Naeemi [20, 21] have investigated such functions and given characterizations that suggest a more general result is possible. Another potential application is scoring rules for alternate representations of uncertainty, several of which result in a decision maker optimizing a convex function [55].

In one sense getting such a characterization is straightforward. In the affine case we want $\mathfrak{A}(t')(t)$ to be an affine function such that $\mathfrak{A}(t')(t) \leq G(t)$ and $\mathfrak{A}(t')(t') = G(t')$. Since we have fixed its value at a point, the only freedom we have is in the linear part of the function, and being a subgradient is exactly the definition of such a linear function. So while our characterization of affine scores is in some sense vacuous, it is also powerful in that it allows us to make use of the tools of convex analysis. A similarly vacuous characterization is possible for the convex case: $\mathfrak{A}(t')(t)$ is a convex function such that $\mathfrak{A}(t')(t) \leq G(t)$ and $\mathfrak{A}(t')(t') = G(t')$. The challenge is to find a way to state it that is useful and naturally handles constraints such as those imposed by the form of a utility function.

While in this dissertation we focus on mechanism design, scoring rules, and prediction markets, another interesting direction to pursue is other settings where our results may be applicable. One natural domain that is closely related to scoring rules for properties is the literature on M-estimators in machine learning, statistics and economics. Essentially, this literature takes a loss function (i.e. a scoring rule) and asks what property it elicits. For example, the mean is an M-estimator induced by the squared error loss function. Some work in this literature (e.g. [75]) requires that the loss function satisfy certain conditions, and our results may be useful in characterizing and supplying such loss functions.

Chapter 3

Duality and general property elicitation

In many settings, it is difficult, or even impossible, to have agents report an entire type $t \in \mathcal{T}$. For example, when allocating a divisible good (e.g. corn) where $\mathcal{O} = \mathbb{R}_+$, a mechanism with $\mathcal{T} = (\mathcal{O} \to \mathbb{R})$ requires agents to submit an infinite-dimensional type. Even type spaces which are exponential in size can be problematic from an algorithmic perspective. Moreover, in many situations, the principal is *uninterested* in all but some small asepect of an agent's private type. For example, the information is often to be used to eventually make a specific decision, and hence only the information directly pertaining to the decision is actually needed — why ask for the agent's entire probability distribution of rainfall tomorrow if a principal wanting to choose between {umbrella, no umbrella} would be content with its expected value?

It is therefore natural to consider a model of truthful reporting where agents provide some sort of summary information about their type. Such a model has been studied in the scoring rules literature, where one wishes to elicit some statistic, or *property*, of a distribution, such as the mean or quantile [92, 65, 50]. We follow this line of research, and extend our affine score framework to accept reports from a different (intuitively, much smaller) space than $\mathcal{T}$.

Our results shed light on the structure of properties and their affine scores. As we shall see, property elicitation is intimately connected to a notion of *duality* between the report space and the type space, which stems fundamentally from vector space duality and convex conjugate duality. We formalize and explore this duality, and our results reveal a very rich theory of elicitation.

3.1 Previous literature on properties

The literature on property elicitation lies essentially entirely within the domain of scoring rules and statistics. This is not to say that one cannot find plenty of cases in mechanism design and other elicitation domains where an agent is asked to report something different

than their true type — indeed, such indirect revelation mechanisms $\mathfrak{M}$ are quite common — but in these models it is usually implicit that either (a) a direct revelation mechanism $\mathfrak{M}'$ will be constructed, which requests the agent's full type and provides the best report to $\mathfrak{M}$ on behalf of the agent, or (b) there is no a priori relationship between the "truthful" report and the underlying type. That is to say, while indirect elicitation has been studied outside of the scoring rules literature, it is only the latter that explicitly requests a mapping, usually a statistic, from the underlying type to the desired information.

As we discussed in §1.2.1, the study of indirect elicitation in scoring rules can be traced to Savage, who considered the problem of eliciting expected values of random variables [92]. Osband [79] goes on to provide a rigorous version, generalizing to expected values of functions of the underlying variable. (We will explore expectation elicitation in Chapter 4 as well as §3.4.2 in the present chapter.) More generally, Gneiting and Raftery [51] and Gneiting [50] consider other common statistics as well, such as quantiles, ratios of expectations, and expectiles.

While these and many other examples of specific statistics have appeared in the literature, it was perhaps Lambert et al. [65] who first considered the following general problem: given an outcome space $\mathcal{O}$ and an *arbitrary* map $\Gamma : \Delta(\mathcal{O}) \to \mathbb{R}$, under what circumstances can we construct a proper scoring rule $\mathfrak{S} : \mathbb{R} \times \mathcal{O} \to \mathbb{R}$ for Γ, i.e. where

$$\Gamma(p) \in \operatorname*{argmax}_{r \in \mathbb{R}} \mathbb{E}_{o \sim p}[\mathfrak{S}(r, o)]$$

for every $p \in \Delta(\mathcal{O})$? Moreover, what is the full classification of functions Γ which can be elicited in this way? Lambert et al. [65] make a number of significant contributions towards these goals for the special case of scalar properties, where Γ is real-valued. Lambert and Shoham [66] also characterize elicitable properties Γ which take on finitely many values, showing a connection to *power diagrams* from computational geometry — we will extend this result to our setting in §3.4.1.

In the present chapter we take a similarly general approach to the problem of indirect elicitation. We ask, given an arbitrary multivalued map Γ which specifies the correct report(s) for a given type, whether Γ can be elicited by an affine score, and if so, which scores elicit it. While this may seem too general a problem to consider, as we have made no assumptions on Γ whatsoever, we find that even at such great heights of abstraction, we can employ our tools from Chapter 2 and from convex duality to make precise and useful statements about the nature of indirect (and direct) elicitation.

3.2 Extending our model and characterization

We wish to generalize the notion of an affine score to accept reports from a space $\mathcal{R}$ which is different from $\mathcal{T}$. To even discuss truthfulness in this setting, we need a notion of a truthful report r for a given type t. We encapsulate this notion by a general multivalued map which specifies all (and only) the correct values for t.

Definition 3.1. *Let $\mathcal{R}$ be some given report space. A* property *is a multivalued map $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ which associates a nonempty set of correct report values to each type. We let $\Gamma_r \doteq \{t \in \mathcal{T} \,|\, r \in \Gamma(t)\}$ denote the set of types t corresponding to report value r.*

One can think of Γ_r as the "level set" of Γ corresponding to value r. This concept will be especially useful when we consider finite-valued properties in Section 3.4.1. A natural constraint to impose on these level sets is that they be *non-redundant*, meaning no property value r has a level set entirely contained in another.

Definition 3.2. *Property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ is* redundant *if there exist $r, r' \in \mathcal{R}$ such that $\Gamma_{r'} \subseteq \Gamma_r$. Otherwise, Γ is* non-redundant*.*

The non-redundancy condition is essentially a bookkeeping tool. If one adds report elements r' which are dominated (strictly or otherwise) by another report r, then any time r' would be correct, an agent could safely report r instead. Hence, one could think of imposing this condition then as simply "pre-processing" Γ to remove any dominated reports.

We extend the notion of an affine score to this setting, where the report space is $\mathcal{R}$ instead of $\mathcal{T}$ itself. Note that $\mathcal{A}$ is still a subset of $\mathsf{Aff}(\mathcal{V} \to \overline{\mathbb{R}})$.

Definition 3.3. *An affine score $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$* elicits *a property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ if for all t,*

$$\Gamma(t) = \operatorname*{argsup}_{r \in \mathcal{R}} \mathfrak{A}(r)(t). \tag{3.1}$$

If we merely have $\Gamma(t) \subseteq \operatorname{argsup}_{r \in \mathcal{R}} \mathfrak{A}(r)(t)$, we say $\mathfrak{A}$ weakly elicits *Γ. A property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ is* elicitable *if there exists some affine score $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ eliciting Γ.*

As before, we need a notion of regularity — an affine score $\mathfrak{A}$ is Γ-*regular* if $\mathfrak{A}(r)(t) < \infty$ always and $\mathfrak{A}(r)(t) \in \mathbb{R}$ whenever $r \in \Gamma(t)$. We define Γ-regular linear and affine families similarly.[1]

The simplest way to come up with an elicitable property is to induce one from an affine score. For any $\mathfrak{A} : \mathcal{R} \to \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$, the property

$$\Gamma_{\mathfrak{A}} : t \to \operatorname*{argsup}_{r \in \mathcal{R}} \mathfrak{A}(r)(t) \tag{3.2}$$

is trivially elicited by $\mathfrak{A}$.

Observe also that any affine score $\mathfrak{A}$ eliciting Γ gives rise to a truthful affine score in the original sense — in fact, this is a version of the so-called *revelation principle*. For each t let $r_t \in \Gamma(t)$ be a report choice for t; then the affine score $\mathfrak{A}^{\mathcal{T}}(t')(t) \doteq \mathfrak{A}(r_{t'})(t)$ is truthful. Moreover, by our choices of $\{r_t\}$, we have

$$G(t) \doteq \sup_{t' \in \mathcal{T}} \mathfrak{A}^{\mathcal{T}}(t')(t) = \sup_{r \in \mathcal{R}} \mathfrak{A}(r)(t). \tag{3.3}$$

[1] The family $\{\ell_r \in \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}})\}_{r \in \mathcal{R}}$ is Γ-*regular* if $\ell_r(t) \in \mathbb{R}$ for all $t \in \Gamma_r$, and $\ell_r(t') \in \mathbb{R} \cup \{-\infty\}$ for $t' \neq \Gamma_r$. Likewise for Γ-regular affine functions.

Of course, in general, $\mathfrak{A}^{\mathcal{T}}$ will not be strictly truthful, since by definition, any reports t', t'' with $r_{t'} = r_{t''}$ will have $\mathfrak{A}^{\mathcal{T}}(t') \equiv \mathfrak{A}^{\mathcal{T}}(t'')$. Thus we may think of a property as *refining* the notion of strictness for a truthful affine score. The connection we will draw in Theorem 3.2 is that, in light of (3.3), a property Γ therefore specifies the portions of the domain of $\mathcal{T}$ where G must be "flat".

To get at the connection between properties and "flatness", we start with a technical lemma which shows that having the same subgradient at two different points implies that G is flat in between.

Lemma 3.1. *Let $G : \mathsf{conv}(\mathcal{T}) \to \overline{\mathbb{R}}$ be convex with $G(\mathcal{T}) \subseteq \mathbb{R}$, and let $d \in \partial G_t$ for some $t \in \mathcal{T}$. Then for all $t' \in \mathcal{T}$,*

$$d \in \partial G_{t'} \iff G(t) - G(t') = d(t - t').$$

Proof. First, the forward direction. Applying the subgradient inequality (2.2) at t' for $dG_t = d$ and at t for $dG_{t'} = d$, we have

$$\begin{aligned} G(t') &\geq G(t) + d(t' - t) \\ G(t) &\geq G(t') + d(t - t'), \end{aligned}$$

from which the result follows (as $G(t)$ and $G(t')$ are finite).

For the converse, assume $G(t) = G(t') + d(t - t')$ and let $t'' \in \mathcal{T}$ be arbitrary. Note first that $d(t) \in \mathbb{R}$ as $d \in \partial G_t$, so $d(t') \in \mathbb{R}$ as well. Then using the subgradient inequality 2.2 we have

$$\begin{aligned} G(t') + d(t'' - t') &= G(t') + d(t'' - t) + d(t - t') \\ &= G(t) + d(t'' - t) \\ &\leq G(t''), \end{aligned}$$

□

We are now ready to state our characterization in this setting, which in essence says that eliciting a property Γ is equivalent to eliciting subgradients of a convex function G.

Theorem 3.2. *Let property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ and Γ-regular affine score $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ be given. Then $\mathfrak{A}$ elicits Γ if and only if there exists some convex $G : \mathsf{conv}(\mathcal{T}) \to \overline{\mathbb{R}}$ with $G(\mathcal{T}) \subseteq \mathbb{R}$, and $\varphi : \mathcal{R} \to \partial G$ satisfying $r \in \Gamma(t) \iff \varphi(r) \in \partial G_t$, such that for all $r \in \mathcal{R}$ and $t \in \mathcal{T}$,*

$$\mathfrak{A}(r)(t) \;=\; G(t_r) + \varphi(r)(t - t_r), \tag{3.4}$$

where $\{t_r\}_{r \in \mathcal{R}} \subseteq \mathcal{T}$ satisfies $r' \in \Gamma(t_{r'})$ for all r'.

Proof. For the forward direction, assume that affine score $\mathfrak{A}$ elicits Γ. For each r, we may extend $\mathfrak{A}(r)$ to all $\hat{t} \in \mathsf{conv}(\mathcal{T})$ by linearity as in Theorem 2.1, whence we may define $G(\hat{t}) \doteq \sup_{r \in \mathcal{R}} \mathfrak{A}(r)(\hat{t})$, which is finite for $\hat{t} \in \mathcal{T}$ as $\mathfrak{A}$ is Γ-regular. We wish to show that the choice $\varphi : r \mapsto \mathfrak{A}_\ell(r)$ suffices, where $\mathfrak{A}_\ell$ denotes the linear part of $\mathfrak{A}$.

By standard arguments (see e.g. [88, Prop. 8.12] and [5, §7]), the subgradients to a convex function $F : X \to \overline{\mathbb{R}}$ at $x \in \mathsf{dom}(F)$ are precisely the gradients of its affine supports at x. More formally, let $\mathrm{AS}(F, x) = \{a \in \mathsf{Aff}(X \to \overline{\mathbb{R}}) \,|\, a(x) = F(x),\ \forall x\ F(x) \geq a(x)\}$ be the set of affine supports of F at x. Then for all $x \in \mathsf{dom}(F)$ we have $\partial F_x = \{\partial a \,|\, a \in \mathrm{AS}(F, x)\}$.[2]

By definition of G, we have $\mathfrak{A}(r)(\hat{t}) \geq G(\hat{t})\ \forall \hat{t} \in \mathsf{conv}(\mathcal{T})$. Noting that $\partial \mathfrak{A}(r)_t = \{\mathfrak{A}_\ell(r)\}$, we have for all $t \in \mathcal{T}$ and $r \in \mathcal{R}$,

$$\begin{aligned} r \in \Gamma(t) &\iff \mathfrak{A}(r)(t) = \sup_{r' \in \mathcal{R}} \mathfrak{A}(r')(t) \\ &\iff \mathfrak{A}(r)(t) = G(t) \\ &\iff \mathfrak{A}(r)(t) \in \mathrm{AS}(G, t) \\ &\iff \mathfrak{A}_\ell(r) \in \partial G_t, \end{aligned}$$

where we use elicitation in the first biconditional. Finally, using Γ-regularity and the definition of t_r, we show the form (3.4):

$$G(t_r) + \mathfrak{A}_\ell(r)(t - t_r) = \mathfrak{A}(r)(t_r) + \mathfrak{A}_\ell(r)(t - t_r) = \mathfrak{A}(r)(t).$$

For the converse, let G, φ, and t_r be given, and assume $\mathfrak{A}$ has the form (3.4). First note that by definition of $\{t_r\}_{r \in \mathcal{R}}$, and by assumption on φ, we have $\varphi(r) \in \partial G_{t_r}$ for all r. We claim that G dominates $\mathfrak{A}$, meaning for all $r \in \mathcal{R}$ and all $t \in \mathcal{T}$, $G(t) \geq \mathfrak{A}(r)(t)$; this follows from the definition of $\mathfrak{A}$ and the subgradient inequality (2.2) applied to $\varphi(r)$ at t.

Now by Lemma 3.1, we have

$$\begin{aligned} \varphi(r) \in \partial G_t &\iff G(t) - G(t_r) = \varphi(r)(t - t_r) \\ &\iff G(t) = \mathfrak{A}(r)(t). \end{aligned} \tag{3.5}$$

By definition of a property, we have for all $t \in \mathcal{T}$ that $\Gamma(t) \neq \emptyset$, so by the above $G(t) = \mathfrak{A}(r)(t)$ for some r. Combining this with the fact that G dominates $\mathfrak{A}$, we have $G(t) = \sup_r \mathfrak{A}(r)(t)$ for all $t \in \mathcal{T}$. Putting this together with (3.5) and our assumption that $r \in \Gamma(t) \iff \varphi(r) \in \partial G_t$, we have

$$r \in \Gamma(t) \iff G(t) = \mathfrak{A}(r)(t) \iff r \in \operatorname*{argsup}_{r'} \mathfrak{A}(r')(t). \qquad \square$$

As a corollary, we also obtain a better understanding of weak elicitation, which we will need in the following sections.

Corollary 3.3. *Let property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ and Γ-regular affine score $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ be given. Then $\mathfrak{A}$ weakly elicits Γ if and only if $\mathfrak{A}$ satisfies (3.4) with the weaker condition that $r \in \Gamma(t) \implies \varphi(r) \in \partial G_t$.*

[2] Briefly, the argument is as follows: $\ell \in \partial \mathrm{AS}(F, x)$ implies $\exists c, v\ F(x) = c + \ell(x - v)$ and $\forall x'\ F(x') \geq c + \ell(x' - v)$. Then $F(x) + \ell(x' - x) = c + \ell(x - v) + \ell(x' - x) = c + \ell(x' - v) \leq F(x')$ when $F(x) < \infty$. For the converse, take $c = F(x)$ and $v = x$.

Proof. Given any affine score $\mathfrak{A}$, and defining $\Gamma_{\mathfrak{A}}$ as in (3.2), we see that $\mathfrak{A}$ weakly elicits Γ if and only if $\Gamma(t) \subseteq \Gamma_{\mathfrak{A}}(t)$ for all t. Now let $\mathfrak{A}$ weakly elicit Γ. As $\mathfrak{A}$ trivially elicits $\Gamma_{\mathfrak{A}}$, we apply Theorem 3.2 and now have in particular $r \in \Gamma(t) \implies r \in \Gamma_{\mathfrak{A}}(t) \implies \varphi(r) \in \partial G_t$. For the converse, simply define $\Gamma_{\mathfrak{A}}(t) = \{r \in \mathcal{R} \mid \varphi(r) \in \partial G_t\}$. By Theorem 3.2, $\mathfrak{A}$ elicits $\Gamma_{\mathfrak{A}}$, and by assumption we have $\Gamma(t) \subseteq \Gamma_{\mathfrak{A}}(t)$ for all t. □

Returning to the above discussion, by focusing on $\mathfrak{A}^{\mathcal{T}}$ instead of $\mathfrak{A}$, we now see how properties are essentially a refinement of strictness. Up to a remapping of $\mathcal{R}$, $\mathfrak{A}^{\mathcal{T}}$ is strictly truthful if and only if $\mathfrak{A}$ elicits $\Gamma : t \mapsto \{t\}$, and $\mathfrak{A}^{\mathcal{T}}$ is truthful if and only if $\mathfrak{A}$ elicits some Γ such that $t \in \Gamma(t)$ for all $\mathcal{T}$. In fact, one can can formalize this using Corollary 3.3: a affine score $\mathfrak{A}$ is truthful if and only if it weakly elicits $\Gamma : t \mapsto \{t\}$. Hence, Theorem 3.2 and Corollary 3.3 are actually a generalization of our main characterization, Theorem 2.1.

This new characterization sheds new light on the structure of elicitable properties. In the scoring rules literature, it is common to assume strong conditions on Γ and $\mathcal{R}$, such as Γ being a function rather than a multivalued map, and Γ being linear [2] or real-valued [65]; we will address these specific cases in §3.5 and §3.4.2. In contrast, Theorem 3.2 allows for an extremely general Γ and $\mathcal{R}$. While our resulting characterization is not as concrete or constructive as those from the scoring rules literature (e.g. the set of elicitable properties is only vacuously characterized), we nonetheless have a very powerful statement which may be of use in particular classes of properties. We illustrate the power of this characterization in the following sections.

3.2.1 The structure of properties

As we mention above, one may interpret Theorem 3.2 as saying that properties are essentially selections of subgradients of a convex function. With a bit more work we can formalize this statement.

Lemma 3.4. *Let Γ, G, and φ be as in Theorem 3.2. If additionally φ is injective, then the condition*

$$\forall r \; \forall t \;\; r \in \Gamma(t) \iff \varphi(r) \in \partial G_t \tag{3.6}$$

is equivalent to

$$\exists \mathcal{D} \subseteq \partial G \;\; s.t. \;\; \forall t \; \varphi(\Gamma(t)) = \partial G_t \cap \mathcal{D}. \tag{3.7}$$

Proof. It is easy to see that the first condition (3.6) is equivalent to $\forall t \; \Gamma(t) = \{r : \varphi(r) \in \partial G_t\}$. Letting $\mathcal{D} = \varphi(\mathcal{R}) \subseteq \partial G$, we then easily have

$$\varphi(\Gamma(t)) = \{\varphi(r) : \varphi(r) \in \partial G_t\} = \partial G_t \cap \mathcal{D}.$$

For the converse, we simply use injectivity of φ and the fact that $\varphi(\mathcal{R}) \subseteq \mathcal{D}$ by definition. □

Furthermore, it is easy to see that this φ must be injective if Γ is to be non-redundant (see Definition 3.2).

Lemma 3.5. *Let Γ, G, φ as in Theorem 3.2. If φ is not injective, Γ is redundant.*

Proof. Suppose $\varphi(r) = \varphi(r')$ for $r \neq r'$. By the condition (3.6), for any t we have $r \in \Gamma(t) \iff \varphi(r) \in \partial G_t \iff \varphi(r') \in \partial G_t \iff r' \in \Gamma(t)$. But then $\Gamma_r = \Gamma_{r'}$, and Γ is redundant. □

In fact, as we will see in §3.3.2, this is a consequence of the constant term in an affine score being a function of the linear term only (specifically, the conjugate dual, G^*). Note that the converse does not hold: a property can be redundant but still have unique subgradients for each report, as the following example illustrates.

Example 3.1. *Let $\mathcal{T} = \mathbb{R}$, $\mathcal{R} = (-\infty, -1] \cup \{0\} \cup [1, \infty)$, and define:*

$$\Gamma(t) = \begin{cases} \{t-1\} & \text{if } t < 0 \\ \{-1, 0, 1\} & \text{if } t = 0 \\ \{t+1\} & \text{if } t > 0 \end{cases}. \tag{3.8}$$

This Γ is redundant, since $\Gamma_0 = \Gamma_{-1} = \Gamma_1 = \{0\}$. As we shall see, however, Γ is elicitable; take

$$\mathfrak{A}(r)(t) = \begin{cases} 0 & \text{if } r = 0 \\ rt - \frac{1}{2}(|r| - 1)^2 & \text{o.w.} \end{cases}. \tag{3.9}$$

Then when $t = 0$, we easily see that $\operatorname{argsup}_{r \in \mathcal{R}} \mathfrak{A}(r)(t) = \{-1, 0, 1\}$. For $t > 0$, by symmetry $r \geq 0$ in the argsup*, and by simple calculus we have $\operatorname{argsup}_{r \in \mathcal{R}} \mathfrak{A}(r)(t) = \{t + 1\}$. Similarly for $t < 0$. Hence, we have shown $\Gamma(t) = \operatorname{argsup}_{r \in \mathcal{R}} \mathfrak{A}(r)(t)$ for all t, as desired.*

Where did this $\mathfrak{A}$ come from? We simply applied Theorem 3.2 with $G(t) = |t| + t^2/2$, representatives $t_r = r - \operatorname{sgn}(r)$, and φ simply being the identity[3]; see Figure 3.1. One can check that $r \in \Gamma(t_r)$ always, that $\varphi(r) \in \partial G_t \iff r \in \Gamma(t)$, and that

$$\begin{aligned} \mathfrak{A}(r)(t) &= G(t_r) + \varphi(r)(t - t_r) \\ &= |r - \operatorname{sgn}(r)| + (r - \operatorname{sgn}(r))^2 + r(t - r + \operatorname{sgn}(r)) \end{aligned}$$

indeed matches (3.8).

Finally, observe that we could also elicit Γ with $G(t) = t^2/2$, $t_r = \varphi(r) = r - \operatorname{sgn}(r)$, yielding:

$$\mathfrak{A}(r)(t) = t_r^2/2 + t_r(t - t_r) = \frac{1}{2}t^2 - \frac{1}{2}(t - t_r)^2. \tag{3.10}$$

Combining Lemmas 3.4 and 3.5 above with Theorem 3.2, we now have the following theorem, which explicitly shows how an elicitable Γ is simply a remapping of (a selection of) subgradients of some convex function.

[3]Strictly speaking, we should write $\varphi(r) \doteq (t \mapsto r \cdot t)$.

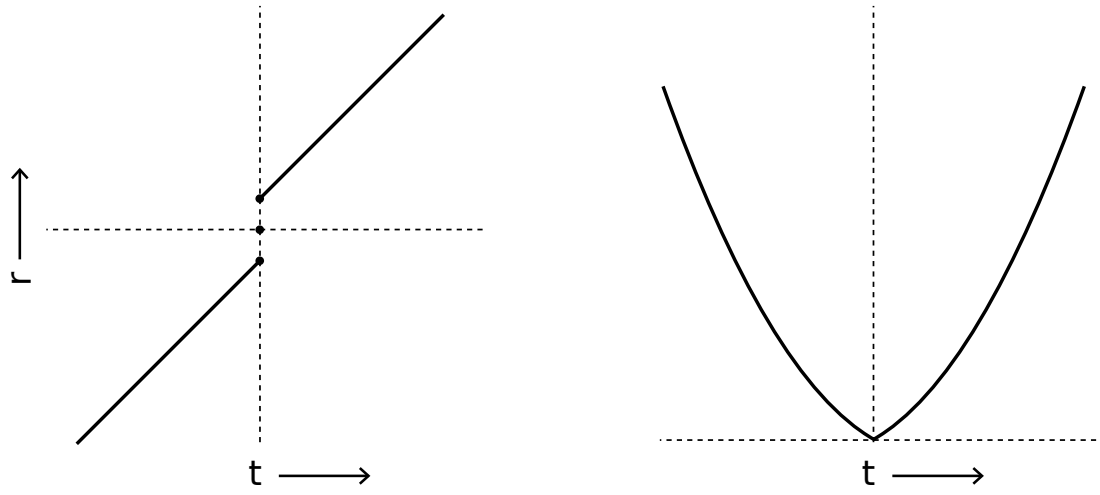


Figure 3.1: The Γ (left) and G (right) from Example 3.1.

Theorem 3.6. *Non-redundant $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ is elicitable if and only if exists there some convex $G : \mathsf{conv}(\mathcal{T}) \to \overline{\mathbb{R}}$ with $G(\mathcal{T}) \subseteq \mathbb{R}$, some $\mathcal{D} \subseteq \partial G$, and some invertible $\varphi : \mathcal{R} \to \mathcal{D}$ such that $\Gamma(t) = \varphi^{-1}(\mathcal{D} \cap \partial G_t)$.*

Note that as discussed earlier the assumption of non-redundancy here is merely a book-keeping concern — of course one can arbitrarily add report elements which are dominated by another (though possibly not strictly), and these cannot be recovered just given the consumer surplus function G. One could even add several copies of the same report, or equivalently, add redundant subgradients at the same point. By imposing non-redundancy we are simply "pre-processing" Γ to remove dominated reports.

A very important question, and one which would give stronger characterizations, is the following:

Question 3.1. *Given non-redundant elicitable Γ, what are all pairs $G, \mathcal{D}$ such that there exists some bijection φ satisfying Theorem 3.6?*

In essence, this question is getting at the fundamental structure of subgradient level sets for convex functions. In §3.4.1 we will see that there is a lot of structure in the *finite* case, where $|\mathcal{R}| < \infty$. In the general case, certainly performing a homothet of the subgradients of G (i.e. scaling G and adding a linear term), will preserve the elicitation structure. However, surely more can be done — the property in Example 3.1 was initially elicited with consumer surplus $G(t) = |t| + t^2/2$, and later with $G(t) = t^2/2$ as well, which is not a homothet transformation.

To illustrate the power of our characterizations, we show some characteristics of general elicitable properties.

Convexity

A well-known property of subgradient mappings is that their level sets are convex.

Proposition 3.7. *For any convex function G, the set $\partial G^{-1}(d) \doteq \{x \in \mathsf{dom}(G) : d \in \partial G_x\}$ is convex.*

Proof. Let $x, x' \in \partial G^{-1}(d)$; then one easily shows (cf. Lemma 3.1) that $G(x) - G(x') = d(x - x')$. Now let $\hat{x} = \alpha x + (1-\alpha)x'$; we have,

$$\begin{aligned} G(\hat{x}) &\leq \alpha G(x) + (1-\alpha)G(x') && (3.11)\\ &= \alpha(G(x) - G(x')) + G(x') \\ &= \alpha d(x - x') + G(x') \\ &= d(\hat{x} - x') + G(x') && (3.12)\\ &\leq G(\hat{x}), && (3.13) \end{aligned}$$

where we applied convexity of G in (3.11) and the subgradient inequality for d at x' in (3.13). Hence, by eq. (3.12) we have shown $G(\hat{x}) - G(x') = d(\hat{x} - x')$, so by Lemma 3.1, $d \in \partial G_{\hat{x}}$. □

In light of our characterizations, this fact about convex functions immediately applies to elicitable properties:

Corollary 3.8. *If $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ is elicitable, then Γ_r is convex for all r.*

To see this, just note that $\varphi(r) \in \partial G_t \cap \partial G_{t'}$ implies that $\varphi(r) \in \partial G_{\hat{t}}$ for all $\hat{t} = \alpha t + (1-\alpha)t'$. Corollary 3.8 generalizes a similar result for scoring rules from Lambert et al. [65] to affine scores.

Cardinality

Combining Theorem 3.2 with the fact that finite-dimensional convex functions are differentiable almost everywhere (cf. [5, Thm 7.26]) yields the following corollary, which shows that elicitable properties have unique values almost everywhere. Note that in some cases this holds in infinite-dimensional vector spaces as well; see e.g. [24, p. 195] and [5, p. 274].

Corollary 3.9. *Let $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ be an elicitable property with $\mathcal{T} \subseteq \mathcal{V} = \mathbb{R}^n$. If $\mathcal{T}$ is of positive measure in $\mathsf{conv}(\mathcal{T})$, and Γ is non-redundant, then $|\Gamma(t)| = 1$ almost everywhere.*

3.3 Duality in elicitation

In the previous section, we took Theorem 3.2 and delved deeper into its conditions, showing that in a strong sense Γ is like a subgradient mapping of a convex function. We now delve even deeper, to the very core of elicitation, and out the other side. The first step is removing the word "like" from the sentence above — we must look at properties which *are* subgradient mappings.

3.3.1 Direct elicitation

Now that we have formalized the relationship between the report space and subgradients of convex functions, we can see what the "canonical" properties look like: those which are (subsets of) subgradient mappings of a convex function. For these properties, we can talk about *direct elicitation*, which roughly speaking means removing the "middle man" φ between $\mathcal{R}$ and ∂G. In fact, for such "canonical" properties, we can even talk about a convex function *itself* eliciting Γ.

Definition 3.4. *A property* $\Gamma : \mathcal{T} \rightrightarrows \mathcal{D}$*, where* $\mathcal{T} \subseteq \mathcal{V}$ *and* $\mathcal{D} \subseteq \mathcal{V}^* \doteq \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}})$*, is* directly elicitable *if there exists* $G : \mathsf{conv}(\mathcal{T}) \to \overline{\mathbb{R}}$ *convex with* $G(\mathcal{T}) \subseteq \mathbb{R}$ *such that* $\Gamma(t) \subseteq \partial G_t$*. In this case we say* G directly elicits*, or just* elicits*,* Γ*.*

In other words, G elicits $\Gamma : \mathcal{T} \rightrightarrows \mathcal{D}$ if the φ in Theorems 3.2 and 3.6 is the identity. Of course, it remains to be shown that there exists an affine score eliciting such a property, but the proof is trivial.

Proposition 3.10. *Directly elicitable properties are elicitable.*

Proof. Let $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ and $G : \mathcal{T} \to \overline{\mathbb{R}}$ convex with $G(\mathcal{T}) \subseteq \mathbb{R}$ be given such that $\Gamma(t) \subseteq \partial G_t$. Then taking $\mathcal{D} = \mathcal{R}$ and $\varphi = \mathrm{id}_{\mathcal{D}}$, we have by Theorem 3.6 that Γ is elicitable. □

Note that this direct elicitability in no way necessary for elicitability, since the report space is not required to have any intrinsic meaning. For example, one can take $\Gamma(t) \doteq -\partial G_t$ for some G, which will not be *directly* elicitable, but still elicitable with $\varphi(r) = -r$ and G.

The notion of direct elicitation is often useful for generating intuitive examples, since the report space itself has meaning. In fact, given any convex function G, the property $\Gamma(t) = \partial G_t$ is directly elicitable by G. This is in fact how Example 3.1 was generated, specifically equation (3.8), though at $t = 0$ we selected $\{-1, 0, 1\}$ instead of the full subgradient set $\partial G_0 = [-1, 1]$.

As a final remark, we note a few observations about direct elicitation. One first notices that the G eliciting some Γ is not unique, as $G' \doteq G + c$ will also elicit Γ for any constant c. But these are the *only* convex functions directly eliciting Γ. Moreover, recovering such a G from Γ is easy: simply integrate (a selection of) Γ to obtain G. Testing whether Γ is directly elicitable is less straight-forward, but there are a variety of monotonicity conditions addressing this issue as well; see Section 5.1.1.

3.3.2 Report duality

We are now ready to hold up a mirror to properties and their scores, and see what we find. That is, we introduce notions of duality. As we will see, there are actually *two* mirrors, yielding four combinations of dualities (see Table 3.1). In this subsection we will explore the first, flipping the report from the type to the dual type.

For now, we will take our dual vector space to be all linear functions from $\mathcal{V}$ to $\mathbb{R}$ (*not* $\overline{\mathbb{R}}$ as above), but in §3.3.3 we will further require $d \mapsto d(v)$ to be linear for all $v \in \mathcal{V}$.[4] We begin with the fundamental object of convex duality, the convex conjugate.

Definition 3.5. *Let* $\mathcal{V}^* \doteq \mathsf{Lin}(\mathcal{V} \to \mathbb{R})$. *The* convex conjugate *of* $G : \mathcal{V} \to \overline{\mathbb{R}}$, *denoted* $G^* : \mathcal{V}^* \to \overline{\mathbb{R}}$, *is given by*

$$G^*(d) = \sup_{v \in \mathcal{V}} \; d(v) - G(v). \tag{3.14}$$

The power of the conjugate, even in this very general setting, is apparent after the following lemma, which says roughly that the convex conjugate "encodes" the subgradients of G. This is a classical result in convex analysis (cf. [95, Thm E.1.4.1]), and we will use it throughout the next two chapters.

Lemma 3.11. *Let* $G : \mathcal{V} \to \overline{\mathbb{R}}$ *be convex. Then for all* $v \in \mathcal{V}, d \in \mathcal{V}^*$,

$$G^*(d) = d(v) - G(v) \iff d \in \partial G_v.$$

Proof. We can simply break down the conditions step by step:

$$\begin{aligned} G^*(d) = d(v) - G(v) &\iff v \in \operatorname*{argsup}_{v' \in \mathcal{V}} \; d(v') - G(v') \\ &\iff \forall v' \in \mathcal{V},\; d(v) - G(v) \geq d(v') - G(v') \\ &\iff \forall v' \in \mathcal{V},\; G(v') \geq G(v) + d(v' - v), \end{aligned}$$

where in the last step we merely negated and added $d(v') \in \mathbb{R}$ to both sides. □

Lemmas 3.4, 3.5, and 3.11 let us further simplify Theorem 3.2, as follows.

Theorem 3.12. *Let non-redundant property* $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ *and* Γ*-regular affine score* $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ *be given. Then* $\mathfrak{A}$ *elicits* Γ *if and only if there exists some convex* $G : \mathsf{conv}(\mathcal{T}) \to \mathbb{R}$, *and bijective* $\varphi : \mathcal{R} \to \mathcal{D}$ *with* $\mathcal{D} \subseteq \partial G$ *satisfying* $\varphi(\Gamma(t)) \subseteq \partial G_t$, *such that for all* $r \in \mathcal{R}$ *and* $t \in \mathcal{T}$,

$$\mathfrak{A}(r)(t) \;=\; \varphi(r)(t) - G^*(\varphi(r)). \tag{3.15}$$

We can now see that there are in fact "canonical scores" as well: every directly elicitable $\Gamma : \mathcal{T} \rightrightarrows \mathcal{D}$ is elicited by some $\mathfrak{A}^{\mathcal{D}}(d)(t) = \langle t, d \rangle - G^*(d)$, and moreover any $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ with $\mathcal{D}$ and φ is elicited (only) by scores $\mathfrak{A}(r)(t) = \mathfrak{A}^{\mathcal{D}}(\varphi(r))(t)$. In other words, properties are in a very literal sense just subsets of subderivative mappings, up to some bijection (or *link function*) taking them to some other report space $\mathcal{R}$.

[4] When the dual space can take on infinite values, the conjugate is not always well-defined, as values of the form $\infty - \infty$ are encountered.

Dual-report mechanisms and the taxation principle. The notion of a dual-report mechanism is already well-known as a consequence of the *taxation principle* — instead of asking the agent for her type, one could simply ask the agent directly for the desired allocation, posting a menu prices (or "taxes") for each. This is without loss of generality because a mechanism's prices cannot depend on the agent's type except through the chosen allocation. In our notation, each allocation d is listed with its price $G^*(d)$. It is worth noting however that this is not always identical to the original mechanism. Specifically, while the equilibrium payoffs for the posted-price mechanism $\mathfrak{A}(d)(t)$ are the same as those of the direct revelation mechanism $\mathfrak{A}(t')(t)$, the *off-equilibrium* payoffs need not be equivalent, as the posted-price mechanism may allow reports $d \in \partial G_t$ which are not $dG_{t'}$ for any t'. In other words, because the primal-report (i.e., direct) mechanism must choose a single subgradient dG_t for every point, if $\{dG_t\}_{\mathcal{T}} \subsetneq \partial G = \mathcal{D}$, the dual-report mechanism can be strictly more expressive. We can see this discrepancy in Example 3.1, since there are three report choices $r \in \{-1, 0, 1\}$ that could correspond to type $t' = 0$, and each yields a different affine function of t not expressible by any other type t'.

Dual-report scoring rules and prediction markets. As we will see in §4.2, the notion of report duality exactly captures the relationship between scoring rules and prediction markets. Here the scoring rules have the primal report space, and prediction markets the dual, where the optimal share bundle is essentially a subgradient of the scoring rule at the trader's belief. There we will further discuss conditions for which the duality can be run in reverse without loss of generality, but as mentioned above about mechanisms, in general the "menu" format (dual report) of an affine score can be strictly more expressive than the type format (primal report).

3.3.3 Type duality and the duality quadrangle

Beyond dual report spaces, we now explore a less familiar notion of duality, defining dual *properties* and their scores, where we completely swap the roles of types and reports. This is the second "mirror," and with both in hand now we have a full four combinations of dual report and type, which we call the duality quadrangle; see Table 3.1.

To start, we need a dual vector space with more structure than simply $\mathsf{Lin}(\mathcal{V} \to \mathbb{R})$. For this we use the notion of a *dual pair*, which is a standard setting for convex analysis in infinite-dimensional spaces.

Definition 3.6 ([5, §5.14])**.** *A pair of topological vector spaces $(\mathcal{V}, \mathcal{V}^*)$ is a* dual pair *if it is equipped with a bilinear form $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V}^* \to \mathbb{R}$ which separates points, in the sense that $\forall v^* \ \langle v, \cdot \rangle \equiv 0$ implies $v = 0$ and $\forall v \ \langle \cdot, v^* \rangle \equiv 0$ implies $v^* = 0$.*

Note that the above assumption that $(\mathcal{V}, \mathcal{V}^*)$ is a dual pair implies in particular that for all $v^* \in \mathcal{V}^*$, the map $v^* \mapsto \langle v, v^* \rangle$ is linear. This will be crucial when interpreting $\mathcal{R} \subseteq \mathcal{V}^*$ as the type space, since affine scores must be affine in the type. Note that as $\mathbb{R}$ is Hausdorff, $\mathcal{V}$

together with the product topology inherited from the dual pair is also Hausdorff and locally convex; see [5, §7] for details. For the remainder of this section (§3.3) we will assume that we have a dual pair $(\mathcal{V}, \mathcal{V}^*)$.

A subject that will come up many times in this section, and even more in Chapter 4, is the conditions under which we have $G^{**} \doteq (G^*)^* = G$. That is, when is the conjugacy operator an involution? Fortunately, in the setting of dual pairs, this has been thoroughly studied in convex analysis. We state the classic theorem due to Fenchel and Moreau [59, 62, 100], which will be useful throughout.

Definition 3.7. *A function $f : X \to \overline{\mathbb{R}}$ is* lower semi-continuous (l.s.c.) *if for every x_0 in $\mathsf{dom}(f)$ it holds that $\liminf_{x \to x_0} f(x) \geq f(x_0)$.*

Theorem 3.13 (Fenchel–Moreau)**.** *Let X be a Hausdorff locally convex space. For any function $G : X \to \overline{\mathbb{R}}$, it follows that $G = G^{**}$ if and only if one of the following is true*

(i) G is a proper, l.s.c., and convex function

(ii) $G \equiv +\infty$, or

(iii) $G \equiv -\infty$.

The following corollary will prove very helpful in our discussion of type duality below, as well as in Chapter 4. The proof follows from applying Theorem 3.13 (recall that dual pairs are automatically Hausdorff and locally convex), and then Lemma 3.11 twice, once for G and once for G^*.

Corollary 3.14. *If G is convex, proper, and l.s.c., then $v^* \in \partial G_v \iff v \in \partial G^*_{v^*}$.*

We now introduce the concept of a *dual property* Γ^*, which essentially swaps the type and the report. That is, an agent has a "true report" r and $\Gamma^*(r)$ encodes all the "correct types" t. We then go on to show the relationship between the direct elicitability of dual properties. See below for possible interpretations of dual properties.

Definition 3.8. *Let $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ where $\mathcal{R} \subseteq \mathcal{V}^*$. Then the* dual *of Γ, written $\Gamma^* : \mathcal{R} \rightrightarrows \mathcal{T}$, is defined by $\Gamma^* \doteq \Gamma^{-1}$. In other words, Γ^* satisfies $r \in \Gamma(t) \iff t \in \Gamma^*(r)$.*

Theorem 3.15. *For dual pair $(\mathcal{V}, \mathcal{V}^*)$, let $\Gamma : \mathcal{T} \rightrightarrows \mathcal{D}$ be given with $\mathcal{T} \subseteq \mathcal{V}$ and $\mathcal{D} \subseteq \mathcal{V}^*$. Let convex proper and l.s.c. G be given. Then G elicits Γ if and only if G^* elicits Γ^*.*

Proof. We apply Corollary 3.14 to obtain $d \in \partial G_t \iff t \in \partial G^*_d$. If G directly elicits Γ, then we have

$$t \in \Gamma^*(d) \iff d \in \Gamma(t) \iff d \in \partial G_t \iff t \in \partial G^*_d,$$

so G^* directly elicits Γ^*. Clearly the above may be applied in the reverse direction as well, yielding the result. □

Let us look deeper into Theorem 3.15. Note that when G and G^* elicit Γ and Γ^*, respectively, we have by the above discussion that $\mathfrak{A}(d)(t) = \langle t, d\rangle - G^*(d)$ elicits Γ and $\mathfrak{A}^*(t)(d) = \langle t, d\rangle - G(t)$ elicits Γ^*. Moreover, the "consumer surplus" functions of $\mathfrak{A}$ and $\mathfrak{A}^*$ are G and G^*, respectively. This curious relationship, combined with the notion of report duality, can be visualized as shown in Table 3.1. Note that traveling around the table does not necessarily mean arriving at the same choice of G, nor does it imply that $G^{**} = G$. However, when $G^{**} = G$ does hold, the diagram "commutes" in a certain sense.

Report \ Type	Primal	Dual
Primal	$\mathfrak{A}(t')(t) = G(t') + \langle t - t', dG_{t'}\rangle$	$\mathfrak{A}^*(t')(d) = \langle t', d\rangle - G(t')$
Dual	$\mathfrak{A}(d')(t) = \langle t, d'\rangle - G^*(d')$	$\mathfrak{A}^*(d')(d) = G^*(d) + \langle dG^*_{d'}, d - d'\rangle$
	$\sup \mathfrak{A}(\cdot)(t) = G(t)$	$\sup \mathfrak{A}^*(\cdot)(d) = G^*(d)$

Table 3.1: The duality quadrangle.

The implications of these dualities, and in particular of type duality, are not yet fully clear. In the following paragraphs we explore various identities and ideas that naturally arise, but leave the rest to future work.

Identities

Table 3.1 shows that the theory of elicitation inherits a lot of structure from convex duality. Ignoring boundary and regulatory concerns for the moment, we obtain some nice identities:

$$\mathfrak{A}(d)(t) + \mathfrak{A}^*(t)(d) \geq \langle t, d\rangle \tag{3.16}$$
$$\mathfrak{A}(d)(t) - \mathfrak{A}^*(t)(d) = G(t) - G^*(d). \tag{3.17}$$

The first follows from the classic Fenchel-Young inequality [89], the proof of which for G proper follows directly from the definition of the conjugate (Definition 3.5).

Lemma 3.16 (Fenchel-Young inequality)**.** $\forall\, v \in \mathcal{V},\, v^* \in \mathcal{V}^*,\ G(v) + G^*(v^*) \geq \langle v, v^*\rangle$.

The elicitation game

Define a two-player game $M(d,t)$, with row strategies $d \in \mathcal{D}$ and column strategies $t \in \mathcal{T}$, as

$$M(d,t) = \Big(\mathfrak{A}(d)(t),\ \mathfrak{A}^*(t)(d)\Big) = \Big(\ \langle t, d\rangle - G^*(d),\ \langle t, d\rangle - G(t)\Big). \tag{3.18}$$

One could think of the column player as choosing the agent's type, and the row player as choosing the principal's "allocation." Interestingly, this interpretation implies that the row is the agent and the column is the principal (they each choose each other's "type"). Immediately one realizes that the Nash equilibria of this elicitation game M are exactly the set of dual-optimal points (d,t) such that $d \in \partial G_t$ and $t \in \partial G^*_d$. Moreover, the equilibrium payoffs for the Nash (d,t) are $\big(G(t), G^*(d)\big)$.

It is interesting to note the mixed strategies of this game: if $d \sim P_{\mathcal{D}}$ and $t \sim P_{\mathcal{T}}$ independently, the payoffs are

$$M(P_{\mathcal{D}}, P_{\mathcal{T}}) = \Big(\ \langle \bar{t}, \bar{d}\rangle - \mathbb{E}_{P_{\mathcal{D}}}[G^*(d)],\ \langle \bar{t}, \bar{d}\rangle - \mathbb{E}_{P_{\mathcal{T}}}[G(t)]\Big), \tag{3.19}$$

and if $(d,t) \sim P$ is supported only on dual points,

$$\mathbb{E}_P[M(d,t)] = \left(\mathbb{E}_{P|_{\mathcal{T}}}[G(t)],\ \mathbb{E}_{P|_{\mathcal{D}}}[G^*(d)]\right), \tag{3.20}$$

both of which bear resemblance to quantities in Bayesian or randomized mechanism settings.

Score divergences

The score divergence $\mathfrak{A}(t)(t) - \mathfrak{A}(t')(t)$ is a natural notion of "regret" which arises frequently in the scoring rules literature (cf. [51]). Our score divergence, as we define below, is reminiscent of a Bregman divergence, a fact we explore further in §4.2.1.

$$D_{G,dG}(t,t') \doteq \mathfrak{A}(t)(t) - \mathfrak{A}(t')(t) = G(t) - G(t') - \langle t - t', dG_{t'}\rangle\,. \tag{3.21}$$

Note that the first argument to D is the true type, as opposed to our $\mathfrak{A}$ notation. Also note the subscripts to D, which specify both the convex function G and a selection of subgradients. A Bregman divergence requires G to be continuously differentiable, but our definition (3.21) is a natural extension, and has been studied before (cf. [60]). We also use this general notion in Definition 4.1.

Score divergences have many nice properties, like convexity in the first argument, and (directional) differentiability at $t' = t$; see Proposition 3.19. Score divergences also enable reasoning about the magnitude of off-equilibrium payoffs, which can be important in practice, when externalities are often present. For example, Fiat et al. [45] introduce the notion of "strong truthfulness", where the payoff decays as $\|t - t'\|^2$, to design mechanisms that are robust even when agents care about the utility of other agents.

Turning to our various notions of duality, the following are four divergences corresponding to the duality quadrangle, starting in the (primal,primal) setting and moving counterclockwise.

$$D_{G,dG}(t,t') = G(t) - G(t') - \langle t - t', dG_{t'} \rangle \tag{3.22}$$
$$D_{G}(t,d') = G(t) + G^*(d') - \langle t, d' \rangle \tag{3.23}$$
$$D_{G^*,dG^*}(d,d') = G^*(d) - G^*(d') - \langle dG^*_{d'}, d - d' \rangle \tag{3.24}$$
$$D_{G^*}(d,t') = G^*(d) + G(t') - \langle t', d \rangle \,. \tag{3.25}$$

Amazingly, we see that $D_G(t,d) = D_{G^*}(d,t)$ for all t, d (not just dual points). In other words, the loss of reporting d in the primal but having type t is the same as reporting t in the dual but having "type" d. In the context of the elicitation game above, this means that at any pure strategy pair, both players have the same regret, so they both stand to gain the same amount in a best response (though a simultaneous best response will *not* lead to an equilibrium point in general).

Dual scoring rules

The notion of a dual scoring rule is relatively straightforward. The agent is endowed with a private lottery ticket $q : \mathcal{O} \to \mathbb{R}$ which specifies her winnings upon each state of the world $\mathcal{O}$. A principal who would like to know the agent's lottery ticket offers her a scoring mechanism $\mathfrak{S}^*(q', q) = \mathbb{E}_{p(q')}[q]$ which selects a distribution $p(q')$ based on the report q'. That is, the mechanism selects p, the *odds* for the gamble, based on the reported ticket q'. Truthfulness implies that this must be done in a way to maximize the winnings for q'.

To make this slightly more concrete, let us take the dual of the scoring rule

$$\mathfrak{S}(p', o) = G(p') + dG_{p'}(o) - \mathbb{E}_{p'}[dG_{p'}] = dG_{p'}(o) - G^*(dG_{p'}).$$

Then the dual would be

$$\mathfrak{S}^*(q', q) = G^*(q') + dG^*_{q'}(q - q') = \mathbb{E}_{p(q')}[q] - G(p(q')),$$

where $p(q') = dG^*_{q'} \in \mathcal{P}$. Thus, the "menu" format of the affine score would be a list of probability distributions p with prices $G(p)$, where G is the consumer surplus from the original scoring rule. Note that as $p \in \mathcal{P}$, the mechanism cannot enforce strict truthfulness between q and $q + \alpha\mathbb{1}$. To get around this, the principal might add a new outcome with a known payout, i.e. set $\mathcal{O}' = \mathcal{O} \cup \{o'\}$ and offer to pay the agent \$1 if o' materializes, and modify the menu above to contain entries $p' \in \Delta(\mathcal{O}')$. Intuitively, the agent's preferences between the existing lottery ticket and the new payout opportunity reveal the "scale" (in an arithmetic sense) of the agent's original gamble.

As a simple example, one can take $G(p) = -H(p)$ where

$$H(p) = -\int_{\mathcal{O}} p(o) \log p(o) \, d\nu(o)$$

is Shannon entropy (see Chapter 4). In this case, the primal scoring rule is just the logarithmic scoring rule $\mathfrak{S}(p, o) = \log p(o)$. Phrased in dual report format, the dual scoring rule is simply,

$$\mathfrak{S}^*(p, q) = \mathbb{E}_p[q] + H(p) = \mathbb{E}_p[q - \log p]$$

where the optimal odds for q are the familiar exponential weights,

$$p(q)(o) = \nabla(-H)^*(q) = \frac{\exp(q(o))}{\int_{\mathcal{O}} \exp(q(o))d\nu(o)}.$$

Dual mechanisms

Less straightforward than a dual scoring rule, the idea of a *dual mechanism* is nonetheless intriguing, where essentially the roles of the principal and agent in mechanism design are reversed. Here the *agent* devises some score $\mathfrak{A}^* : \mathcal{D} \to \mathsf{Aff}(\mathcal{D} \to \mathbb{R})$ where $\mathcal{D} = \{\mathsf{Eval}_o : o \in \mathcal{O}\}$ and the range of $\mathfrak{A}^*$ is $\{\mathsf{Eval}_o \mapsto \mathsf{Eval}_o[t] : t \in \mathcal{T}\}$. Equivalently, and much more straightforwardly, we can write $\mathfrak{A}^* : \mathcal{O} \to \mathcal{T}$. Thus the principal reports an allocation $o' \in \mathcal{O}$, and the mechanism produces a type $t^{o'} = \mathfrak{A}^*(o')$ based on this report, and finally the principal receives utility $t^{o'}(o)$ for the "true allocation" o.

A possible interpretation is that the principal reports what "object" she has in her hands, the "allocation", and the agent chooses his type; the pricipal's payoff is then the valuation of the agent's chosen type on the true allocation. Note that as always, the roles of revenue and consumer surplus are swapped, in exactly the way one would think — the net utility of the agent in both cases is $G(t)$, and that of the principal in both cases is $G^*(d)$.

3.3.4 Rationalizability

Since Rochet [86] or earlier, it has been observed that the design of truthful mechanisms is similar to the *rationalizability* problem, which is the following. Given some purchase data $D = (x^i, p^i)$, where $x^i \in \mathbb{R}^n$ is the ith bundle of goods and $p^i \in \mathbb{R}^n$ is the ith price vector, is D consistent with a concave utility function (which is quasi-linear in money)? In other words, does there exist concave $u : \mathbb{R}^n \to \mathbb{R}$ such that for all i

$$x^i \in \operatorname*{argmax}_x \{u(x) - p^i \cdot x \mid x \in X(p^i)\}, \tag{3.26}$$

where one usually takes $X(p) = \{x \in \mathbb{R}^n \mid p \cdot x \leq B\}$. We will take $X(p) = X$ for the remainder, though this can perhaps be relaxed.

We can make this connection more formal by swapping the roles of the utility and the cost. That is, letting p be the "type", we immediately see that the net utility is affine in p, so we can view this as an affine score. Specifically, letting $\Gamma(p) \subseteq X$ be the set of bundles that the agent buys given prices p, we have that u rationalizes Γ (which encodes D) if and only if $\mathfrak{A}(x)(p) = u(x) - p \cdot x$ elicits Γ.

Note that if we interpret this this affine score as a mechanism, the valuation vector is $t = -p$, meaning agent's utility is actually $t \cdot x = -p \cdot x$, whereas the payment to the mechanism is $-u(x)$. Thus, the consumer surplus function of the mechanism is the conjugate dual of the (negative) utility function, since $G(t) \doteq \sup_{x \in X}\{t \cdot x - (-u)(x)\} = (-u)^*(t)$.

There is certainly duality involved in this conversion between rationalizability and mechanism design, and it is natural to ask how this fits with the various notions of duality discussed above. In both rationalizability and mechanism design, the agent is given prices/type $p = -t$, and must choose a bundle/report x. In terms of the duality quadrangle in Table 3.1, these settings are still in the left-hand side (primal type), specifically the lower left (dual report), though one could argue that the mechanism is the top left (primal report). Thus, the real duality in rationalizability is in taking an affine score $\mathfrak{A}(d)(t) = d(t) - G^*(d)$ and thinking of the $d(t)$ term as the cost and the $-G^*(d)$ term as the utility, rather than the other way around as is typical for mechanism design.

3.4 Characterizations for special cases

We now examine specific classes of properties, using the additional structure to provide stronger characterizations. As discussed in §3.1, the previous work in these areas lies almost entirely in the scoring rules domain, where a property is often interpreted as a *statistic* of a distribution. Our goal will be to extend these results to our much more general setting, both to uncover their deeper mathematical structure and to allow their extention to other elicitation domains; see for example the discussion on properties in mechanism design in §5.2.

The two classes we consider are *finite-valued* properties, where $\mathcal{R}$ is a finite set of possible reports, and *functional* properties, where the multivalued map $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ is actually a function $\Gamma : \mathcal{T} \to \mathcal{R}$ specifying a unique correct report for each $t \in \mathcal{T}$. The latter case we will further break down into the linear case in §3.4.2, where $\mathcal{R}$ is a vector space and Γ is a linear map, and the general nonlinear case in §3.4.3, where we will be forced to make further differentiability assumptions to make headway.

3.4.1 Finite-valued properties

We first consider the case where $\mathcal{R}$ is a finite set of reports. In the scoring rules literature, Lambert and Shoham [66] view this case as eliciting answers to multiple-choice questions. There are also natural applications to mechanism design, which we mention in §5.2.

Our approach is heavily inspired by Lambert and Shoham [66]. Assume throughout that $\mathcal{R}$ is finite and that $\mathcal{T}$ is a convex subset of a vector space $\mathcal{V}$ endowed with an inner product, so that we may write $\langle t, t' \rangle$ and in particular $\|t\|^2 = \langle t, t \rangle$. In this more geometrical setting, we will use the concept of a power diagram from computational geometry.

Definition 3.9. *Given a set of points $P = \{p_i\}_{i=1}^m \subset \mathcal{V}$, called* sites*, and weights $w \in \mathbb{R}^m$, a* power diagram *$D(P, w)$ is a collection of cells* $\mathrm{cell}(p_i) \subseteq \mathcal{T}$ *defined by*

$$\mathrm{cell}_{P,w}(p_i) = \left\{ t \in \mathcal{T} \;\middle|\; i \in \operatorname*{argmin}_j \left\{ \|p_j - t\|^2 - w_j \right\} \right\}. \tag{3.27}$$

The following result is a generalization of Theorem 4.1 of Lambert, et al. [66], and is essentially a restatement of results due to Aurenhammer [11, 9].

Theorem 3.17. *A property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ for finite $\mathcal{R}$ is elicitable if and only if the level sets $\{\Gamma_r\}_{r \in \mathcal{R}}$ form a power diagram $D(P, w)$.*

Proof. Let us examine the condition that t is an element of $\mathrm{cell}_{P,w}(p_i)$ for some power diagram $D(P, w)$:

$$\begin{aligned} t \in \mathrm{cell}_{P,w}(p_i) &\iff i \in \operatorname*{argmin}_j \left\{ \|p_j - t\|^2 - w_j \right\} \\ &\iff i \in \operatorname*{argmin}_j \left\{ \|p_j\|^2 - 2\langle p_j, t \rangle - w_j \right\}. \end{aligned} \tag{3.28}$$

Note that eq. (3.28) is affine in t. Now given some $D = D(P, w)$ with index set $\mathcal{R}$, we simply let $\mathfrak{A}(r)(t) = 2\langle p_r, t \rangle + w_r - \|p_r\|^2$. By (3.28) we immediately have $r \in \operatorname{argsup}_{r'} \mathfrak{A}(r')(t) \iff t \in \mathrm{cell}_{P,w}(p_r)$, as desired.

Conversely, let an affine score $\mathfrak{A}$ eliciting Γ be given. Note that since we are in an inner product space, we may write $\mathfrak{A}(r)(t) = \langle x_r, t \rangle + c_r$ for $x_r \in \mathcal{V}$ and $c_r \in \mathbb{R}$. Letting $p_r = x_r/2$ and $w_r = \|p_r\|^2 + c_r$, we see by (3.28) again that $\Gamma_r = \mathrm{cell}(p_r)$ of the diagram $D(\{p_r\}, w)$. Hence, Γ is a power diagram. □

We have now seen exactly what kinds of finite-valued properties are elicitable, but how can we elicit them? Or more precisely, as the proof above is constructive enough to give sufficient conditions, what are all ways of eliciting a given power-diagram? In general, it is difficult to provide a "closed form" answer to this question, so we restrict to the *simple* case, where essentially the cells of a power diagram are as constrained as possible.

Definition 3.10 ([10])**.** *A* j-polyhedron *is the intersection of dimension j of a finite number of closed halfspaces of $\mathcal{V}$, where $0 \leq j \leq \dim(\mathcal{V}) < \infty$. A* cell complex *C in $\mathcal{V}$ is a covering of $\mathcal{V}$ by finitely many j-polyhedra, called* j-faces *of C, whose (relative) interiors are disjoint and whose non-empty intersections are faces of C. C is called* simple *if each of its j-faces is in the closure of exactly $(d - j + 1)$ d-faces (cells).*

Theorem 3.18. *Let finite-valued, elicitable, simple property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$ be given. Then there exist points $\{p_r\}_{\mathcal{R}} \subseteq \mathcal{V}$ such that an affine score $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ elicits Γ if and only if there exist $\alpha > 0$, and $p_0 \in \mathcal{V}$ such that*

$$\mathfrak{A}(r)(t) = 2\langle \alpha p_r + p_0, t \rangle - \|\alpha p_r + p_0\|^2 + w_r, \tag{3.29}$$

where the choice $w \in \mathbb{R}^{\mathcal{R}}$ is determined by α and p_0.

Proof. A result of Aurenhammer for simple cell complexes, given in Lemma 1 of [9] and the proof of Lemma 4 of [11], states the following: given sites P and P' and weights w, there exist weights w' such that $D(P', w') = D(P, w)$ if and only if P' is a homothet (translated and positively scaled copy) of P. We simply apply this fact to the proof of Theorem 3.17. □

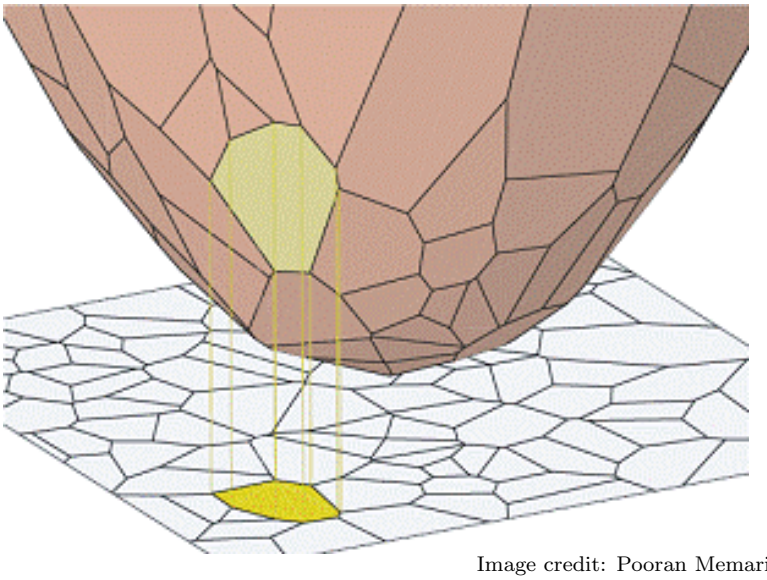


Image credit: Pooran Memari

Figure 3.2: A consumer surplus function G and its corresponding partition of the type space, Γ. The proof of Theorem 3.17 leverages the fundamental relationship between projections of convex functions and power diagrams.

We conclude our exploration of finite-valued properties with a few remarks.

Bregman Voronoi digrams and the role of $\|\cdot\|^2$. The squared norm seems fundamental to our derivation above; let us dig further to see if this is indeed the case. Observe that the form (3.29) is simply

$$\mathfrak{A}(r)(t) = 2\langle t_r, t\rangle - \|t_r\|^2 + w_r,$$

where $t_r = \alpha p_r + p_0$. Consider the case where $w_r = 0$ for all r, which corresponds to Γ being a *Voronoi diagram*. In this case, could think of $\mathfrak{A}$ as being a special case of the "Brier score" $\mathfrak{A}^B(t')(t) = 2\langle t,' t\rangle - \|t'\|^2$, so that $\mathfrak{A}(r)(t) = \mathfrak{A}^B(t_r)(t)$. In other words, we can think of our finite-report case as just restricting the allowed reports in a general direct-revelation affine score. Note that the score divergence for $\mathfrak{A}^B$ is just $D_G(t', t) = \|t' - t\|^2$, where $G(t) = \|t\|^2$ is just the square norm.

This raises the following interesting question: what do we get when we replace $G = \|\cdot\|$ with another convex function on $\mathcal{T}$, and restrict the reports from $\mathcal{T}$ to just a few points $\{t_r\}_{\mathcal{R}}$? That is, take $\mathfrak{A}^G(t')(t) = G(t') - dG_{t'}(t - t')$ and set $\mathfrak{A}(r)(t) = \mathfrak{A}^G(t_r)(t)$. Surely,

for any such G, whatever Γ is elicited by such a modified affine score would have to be a diagram by Theorem 3.17. But then why does the squared norm seem so fundamental?

As it happens, we are touching on precisely the notion of a *Bregman Voronoi diagram*, introduced by Nielsen et al. [23, §4]. There, instead of defining $\text{cell}_i = \{t : i \in \text{argmin}_j \|t_j - t\|\}$, the squared norm is replaced by any Bregman divergence D_G, so that $\text{cell}_i = \{t : i \in \text{argmin}_j D_G(t, t_j)\}$.[5] Our conclusion that such diagrams coincide with power diagrams corresponds to their Theorem 8.

Framed in terms of our report duality from §3.3.2, we can see this yet another way. We can rewrite the Bregman Voronoi cell as

$$\text{cell}_i = \left\{ t : i \in \operatorname*{argmax}_j \; G(t_j) - dG_{t_j}(t - t_j) \right\}. \tag{3.30}$$

By Lemma 3.11, this can in turn be written

$$\text{cell}_i = \left\{ t : i \in \operatorname*{argmax}_j \; \langle \tilde{t}_j, t \rangle - G^*(\tilde{t}_j) \right\}, \tag{3.31}$$

where $\tilde{t}_j = dG_{t_j}$. Hence, for any convex function G, the sites $\{p_j\}$ and weights w of a power diagram corresponding to the D_G Bregman Voronoi diagram with sites $\{t_j\}$ are given by $p_j = \frac{1}{2} dG_{t_j}$ and $w_j = \frac{1}{4}\|dG_{t_j}\|^2 - G^*(dG_{t_j})$.

Degrees of freedom. It is interesting to ask what the *degree of freedom* is when choosing an affine score to elicit a given property Γ. Note that we have a trivial upper bound of $(d+1) \cdot |\mathcal{R}|$, where $d = \dim(\mathcal{V})$, since one at most chooses a site and weight for each cell. Theorem 3.18 makes it clear that in the *simple* case, the degree of freedom is actually bounded by $d+1$. In other words, one can specify the affine score with no more than $d+1$ real numbers, no matter the size of $\mathcal{R}$.[6]

What about outside the simple case — are there cases when we are more or less restricted? In fact, it is easy to come up with examples where we have much more flexibility.

Example 3.2. *The simplest example of a property whose degree of freedom scales with $\mathcal{R}$ is the "collinear" property, $\Gamma(t) = \lfloor \langle t, x \rangle \rfloor$ for some fixed x, where $\lfloor y \rfloor$ denotes the greatest integer less than or equal to y (the floor function). In a single dimension, this becomes simply $\Gamma(t) = \lfloor t \rfloor$; let us take $\mathcal{T} = [1, N+1)$ and $\mathcal{R} = [N] = \{1, 2, \ldots, N\}$. To elicit Γ, we must choose $\mathfrak{A}(r)(t) = \alpha_r t + c_r$ to satisfy the constraints $\alpha_{r+1}(r+1) + c_{r+1} = \alpha_r r + c_r$ and $\alpha_{r+1} \geq \alpha_r$ for all $r \in \mathbb{N}$. Picking an arbitrary α_1 and c_1, we see that this still leaves one degree of freedom between α_2 and c_2, and so on each time we make a choice. Thus we may choose $\{\alpha_r\}_r$ to be any increasing sequence, and the initial offset c_1, for a total of $N+1 = |\mathcal{R}|+1$ degrees of freedom. It is easy to see that in d dimensions, this example gives $|\mathcal{R}| + d$; pick an initial slope and offset ($d+1$) and slope for each additional hyperplane (1 each, for a total of $|\mathcal{R}| - 1$, since the boundaries must be maintained).*

[5] In [23], three types of diagrams are introduced; here we refer to the first type.

[6] The upper bound comes from the fact that two representations may give equivalent scores.

In the context of scoring rules, Lambert [63] notes that given a finite property Γ, there exists some set of *base scores* $\{b_i\}$ such that any scoring rule which is truthful for Γ can be written as a linear combination $\sum \alpha_i b_i$. It is not shown, however, how many base scores there are for a given Γ, and hence the degree of freedom for specifying a score is unclear. We conjecture that the collinear case described above provides the most flexibility, so that the degree of freedom for any Γ is bounded by $d + |\mathcal{R}|$.

Computing elicitability. It is natural to ask, given a property Γ, can we determine whether Γ is elicitable in polynomial time? By Theorem 3.17, we need only test whether the cells $C = \{\Gamma_r\}_{r \in \mathcal{R}}$ form a power diagram. For the *simple* case, Aurenhammer gives an algorithm for this task, given as "Algorithm Orthogonal Dual" in §2.2 of [10] and comments thereafter. The orthogonal dual algorithm assumes that the cells are stored in an *incidence lattice*, with nodes for each face of C, and edges when faces are incident (a j-dimensional face which contains a $(j-1)$-dimensional face). The runtime of the algorithm is $O(m)$, where m is the number of facets (faces of dimension $d-1$).

More generally, Rybnikov in [90, § 12] presents a polynomial-time algorithm which can detect power diagrams in the general case. His work extends to manifolds even beyond convex polytopes (the projections of which yield power diagrams — see Figure 3.2). We will apply this algorithm to mechanism design in § 5.2.

3.4.2 Linear properties

A natural class of properties are those which specify a unique correct report for each type. In other words, properties Γ which are *functions* from $\mathcal{T}$ to $\mathcal{R}$. This case has been studied extensively in the scoring rules literature, and we explore how to extend these results to our setting. In particular, this section focuses on the case that $\Gamma : \mathcal{T} \to \mathcal{R}$ is a *linear* function, in the algebraic sense. That is, for all $\alpha \in \mathbb{R}$ and all $t_1, t_2 \in \mathcal{T}$ such that $\alpha t_1 + t_2 \in \mathcal{T}$, we have $\Gamma(\alpha t_1 + t_2) = \alpha \Gamma(t_1) + \Gamma(t_2)$.

The setting where types are distributions $\mathcal{T} \subseteq \Delta(\mathcal{O})$ has been studied in both machine learning and in the scoring rules literature. In that case, Γ can be thought of as the mean of a random variable. We will address this case in much greater detail in Chapter 4, and discuss previous work in § 4.1.2. We find that even in the distribution case, our characterization is more general than those currently in the literature due to our lack of differentiability requirements or other major technical conditions.

Here we merely assume $\mathcal{R}$ is itself a subset of a topological vector space, and Γ is continuous (and linear). Since Γ is a function, throughout this subsection (and the next) we will use the notation $\Gamma : \mathcal{T} \to \mathcal{R}$ and write $r = \Gamma(t)$. We will only consider real-valued affine scores, i.e. $\mathcal{A} = \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$, though in principle this restriction could be relaxed to the extended reals as above. We will also assume throughout that $\mathcal{T}$ and $\mathcal{R}$ are convex.

As before, our goal is to characterize all elicitable Γ in this setting, and all affine scores (weakly) eliciting a given Γ. The first question turns out to be quite simple: *any* linear map

Γ is elicitable, by taking the following affine score:

$$\mathfrak{A}(r)(t) = G(r) + dG_r(\Gamma(t) - r), \tag{3.32}$$

where $G : \mathcal{R} \to \mathbb{R}$ is convex. To see this, consider $\hat{r} = \Gamma(t)$ to be the type, and invoke Theorem 2.1. We thus turn to the second question: what forms of $\mathfrak{A}$ other than (3.32) weakly elicit Γ?

To answer this question, we will capitalize on a subtle but important property of truthful affine scores: locally around a point t, the score $\mathfrak{A}(\cdot)(t)$ is "smooth." That is, the optimization for an agent, $\sup_{t'} \mathfrak{A}(t')(t)$, is well-behaved for t' sufficiently close to t. To formalize this statement, we define the directional derivative.

Definition 3.11. *Let $f : \mathcal{T} \to \overline{\mathbb{R}}$, and for all $t \in \mathcal{T}$ and $v \in \mathcal{V}$ define*

$$f(t; v) = \lim_{\epsilon \downarrow 0} \frac{f(t + \epsilon v) - f(t)}{\epsilon} \tag{3.33}$$

to be the directional derivative *of f at t in direction v.*

We can now be more precise in describing the behavior of an affine score near the "diagonal": the directional derivative $\mathfrak{A}(t; t' - t)(t)$ is 0. Note that our compact notation here is just stating $f(t; t' - t) = 0$ for $f(\cdot) = \mathfrak{A}(\cdot)(t)$.

Proposition 3.19. *Let $\mathfrak{A} : \mathcal{T} \to \mathcal{A}$ be a truthful affine score, where $\mathcal{T}$ is convex and $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$. Then for all $t, t' \in \mathcal{T}$ the directional derivative $\mathfrak{A}(t; t' - t)(t)$ exists and is equal to 0.*

Proof. By Theorem 2.1, we have some convex function $G : \mathcal{T} \to \mathbb{R}$, and some selection of subgradients $\{dG_t\}_{t \in \mathcal{T}}$, such that

$$\mathfrak{A}(t')(t) = G(t') + dG_{t'}(t - t'). \tag{3.34}$$

Fix $t, t' \in \mathcal{T}$, and define $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = G(t + x(t' - t))$. Then we have $\mathfrak{A}(t + x(t' - t))(t) = g(x) - dg_x(x)$ where one can check that $dg_x \doteq dG_{t+x(t'-t)}$ is indeed a subgradient to g. Now we have

$$\begin{aligned} \mathfrak{A}(t; t' - t)(t) &= \lim_{\epsilon \downarrow 0} \frac{\mathfrak{A}(t + \epsilon(t' - t))(t) - \mathfrak{A}(t)(t)}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{g(\epsilon) - dg_\epsilon(\epsilon) - g(0)}{\epsilon} \\ &= g'_+(0) - \lim_{\epsilon \downarrow 0} dg_\epsilon, \end{aligned} \tag{3.35}$$

where $g'_+(x)$ is the right derivative of g at x. It is clear that $dg_x \leq g'_+(x)$ for all x, and we have from [89, Theorem 24.1] that $g'_+(x) \leq g'_+(x + \epsilon)$ and $\lim_{\epsilon \downarrow 0} g'_+(x + \epsilon) = g'_+(x)$. Combining these we see that $\lim_{\epsilon \downarrow 0} dg_\epsilon = g'_+(\epsilon)$. Hence we have $\mathfrak{A}(t; t' - t)(t) = 0$ as desired. □

We can now use the revelation principle to apply Proposition 3.19 to our linear property setting. Note that linearity of Γ is key here; without it, passing from $\Gamma(t + \epsilon(t' - t))$ to $r + \epsilon(r' - r)$ would introduce error terms which may not be well-behaved in general.

Corollary 3.20. *Let $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ be given which weakly elicits a linear property Γ, where $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$ and $\mathcal{R}$ and $\mathcal{T}$ are convex. Then for all $t \in \mathcal{T}$ and $r = \Gamma(t)$, and all $r' \in \mathcal{R}$, we have $\mathfrak{A}(r; r' - r)(t) = 0$.*

Proof. The affine score $\mathfrak{A}^{\mathcal{T}}(t')(t) \doteq \mathfrak{A}(\Gamma(t'))(t)$ is truthful because $\mathfrak{A}$ is (this is the "revelation principle"), and hence we may apply Proposition 3.19 to $\mathfrak{A}^{\mathcal{T}}$. For any $t' \in \Gamma_{r'}$,

$$\begin{aligned}
0 = \mathfrak{A}^{\mathcal{T}}(t; t' - t)(t) &= \lim_{\epsilon \downarrow 0} \frac{\mathfrak{A}^{\mathcal{T}}(t + \epsilon(t' - t))(t) - \mathfrak{A}^{\mathcal{T}}(t)(t)}{\epsilon} \\
&= \lim_{\epsilon \downarrow 0} \frac{\mathfrak{A}\Big(\Gamma(t + \epsilon(t' - t))\Big)(t) - \mathfrak{A}\Big(\Gamma(t)\Big)(t)}{\epsilon} \\
&= \lim_{\epsilon \downarrow 0} \frac{\mathfrak{A}(r + \epsilon(r' - r))(t) - \mathfrak{A}(r)(t)}{\epsilon} \\
&= \mathfrak{A}(r; r' - r)(t).
\end{aligned}$$

□

We are getting closer to our goal of understanding affine scores for linear properties. Our next step will be to show that we may extend $\mathfrak{A}$ outside of $\mathcal{T}$ along level sets of Γ. Note that α here is not just in the unit interval, but on the whole real line — we are saying that $\mathfrak{A}$ is still truthful even on *linear extensions* of level sets of Γ.

Lemma 3.21. *Let $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ be given which weakly elicits a linear property $\Gamma : \mathcal{T} \rightrightarrows \mathcal{R}$, where $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$. Then $\mathfrak{A}$ weakly elicits Γ on expanded type space $\hat{\mathcal{T}} = \{\alpha t + (1-\alpha)t' : \alpha \in \mathbb{R},\ \exists r \in \mathcal{R}\ t, t' \in \Gamma_r\}$.*

Proof. Let $r, r' \in \mathcal{R}$, $t, t' \in \Gamma_r$, and $\hat{t} = \alpha t + (1 - \alpha)t'$. Note first that by linearity of Γ we have $\Gamma(\hat{t}) = \alpha\Gamma(t) + (1 - \alpha)\Gamma(t') = r$. Thus,

$$\begin{aligned}
\mathfrak{A}(r')(\hat{t}) &= \mathfrak{A}_\ell(r')(\alpha t + (1 - \alpha)t') + \mathfrak{A}_c(r') \\
&= \alpha\mathfrak{A}_\ell(r')(t) + (1 - \alpha)\mathfrak{A}_\ell(r')(t') + \alpha\mathfrak{A}_c(r') + (1 - \alpha)\mathfrak{A}_c(r') \\
&= \alpha\mathfrak{A}_\ell(r)(t) + (1 - \alpha)\mathfrak{A}_\ell(r)(t') + \alpha\mathfrak{A}_c(r) + (1 - \alpha)\mathfrak{A}_c(r) \\
&= \mathfrak{A}_\ell(r)(\hat{t}).
\end{aligned}$$

□

We are now ready to state the crucial lemma of this section, which says that a truthful affine score for a linear property depends on the type only through Γ, modulo some linear term independent of the report.

Lemma 3.22. *Let $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ be given which weakly elicits a linear property Γ, where $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$ and $\mathcal{R}$ and $\mathcal{T}$ are convex. Then there exists $\ell \in \mathsf{Lin}(\mathcal{V} \to \mathbb{R})$ such that for all $r, r' \in \mathsf{relint}(\mathcal{R})$ and all $t_1, t_2 \in \Gamma_r$,*

$$\mathfrak{A}(r')(t_1) - \mathfrak{A}(r')(t_2) = \ell(t_1 - t_2). \tag{3.36}$$

Proof. Let $t_d = t_1 - t_2 \in \mathcal{V}$ such that $\Gamma(t_1) = \Gamma(t_2)$. By linearity of Γ we have $\Gamma(t_d) = 0$, and moreover, $\Gamma(t+t_d) = \Gamma(t)$ for all $t \in \mathcal{T}$. Thus, these "level-set differences" t_d are independent of the Γ-value of the level set; in light of this, we define $\mathcal{T}^D = \{t_d \in (\mathcal{T} - \mathcal{T}) | \Gamma(t_d) = 0\} \subseteq \ker(\Gamma)$.

Now let $t \in \mathsf{relint}(\mathcal{T})$ and $t_d \in \mathcal{T}^D$ be arbitrary. By definition of relint (see §2.2 or [103, pp. 2-3]), we have some $\delta > 0$ such that $t + \delta t_d \in \mathcal{T}$. Note that $\Gamma(t + \delta t_d) = \Gamma(t) + 0 = r$ for some $r \in \mathcal{R}$. Thus, by Lemma 3.21 we have that $\mathfrak{A}$ can be extended to $\alpha(t+\delta t_d)+(1-\alpha)t = t+\alpha\delta t_d$ for all $\alpha \in \mathbb{R}$; that is, the set $\Gamma_r + \mathrm{span}(\mathcal{T}^D)$. Hence, we conclude that $\mathfrak{A}$ weakly elicits Γ on extended type space $\hat{\mathcal{T}} \doteq \mathcal{T} \cup (\mathsf{relint}(\mathcal{T}) + \mathrm{span}(\mathcal{T}^D))$. Note that as $\mathcal{T}$ is convex and $\mathcal{R} = \Gamma(\mathcal{T})$, we have $\mathsf{relint}(\mathcal{R}) = \Gamma(\mathsf{relint}(\mathcal{T}))$ (cf. [88, Prop. 2.44]), and in particular, $\Gamma_r + \mathrm{span}(\mathcal{T}^D) \subseteq \hat{\mathcal{T}}$ for all $r \in \mathsf{relint}(\mathcal{R})$.

We now break $\mathfrak{A}$ into its linear $\mathfrak{A}_\ell$ and constant $\mathfrak{A}_c$ parts, so that $\mathfrak{A}(r)(t) = \mathfrak{A}_\ell(r)(t) + \mathfrak{A}_c(r)$. Observe that for fixed r, we have $\mathfrak{A}(r)(t_1) - \mathfrak{A}(r)(t_2) = \mathfrak{A}_\ell(r)(t_1 - t_2)$, as the constant terms cancel out. We now show that $\mathfrak{A}_\ell(r; r' - r)(t_d) = 0$ for all $r \in \mathsf{relint}(\mathcal{R})$, $r' \in \mathcal{R}$, using Corollary 3.20:

$$\mathfrak{A}_\ell(r; r' - r)(t_d) = \mathfrak{A}(r; r' - r)(t_r + t_d) - \mathfrak{A}(r; r' - r)(t_r) = 0 + 0 = 0$$

for any $t_r \in \Gamma_r$. Note that restriction $r \in \mathsf{relint}(\mathcal{R})$ is necessary to guarantee that $\mathfrak{A}$ is defined (and truthful) for the point $t_r + t_d$.

Finally, as we have now shown that $\mathfrak{A}_\ell(r)(t_d)$ is continuously Gâteaux differentiable (see [5, §7]) at all $r \in \mathsf{relint}(\mathcal{R})$, and all directional derivatives are 0, we conclude that it must be a constant function (in r). Now fixing some $r \in \mathsf{relint}(\mathcal{R})$ and letting $\ell(t_d) \doteq \mathfrak{A}_\ell(r)(t_d)$ for all $t_d \in \mathcal{T}^D$ concludes the proof. □

From Lemma 3.22, the main result of this section now follows. Note that we must be careful about the relative interior of $\mathcal{R}$ — to this end, we say that $\mathfrak{A}$ weakly elicits Γ *on* $\mathcal{T}'$ if $\Gamma(t') \in \mathrm{argsup}_r \mathfrak{A}(r)(t')$ for all $t' \in \mathcal{T}'$. Similarly, we say $\mathfrak{A}$ weakly elicits Γ on $\mathcal{R}'$ if it weakly elicits Γ on $\Gamma^{-1}(\mathcal{R}')$.

Theorem 3.23. *Let linear property $\Gamma : \mathcal{T} \to \mathcal{R}$ and Γ-regular affine score $\mathfrak{A} : \mathcal{R} \to \mathcal{A}$ be given, where $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{V} \to \mathbb{R})$ and $\mathcal{R}$ and $\mathcal{T}$ are convex. Then $\mathfrak{A}$ weakly elicits Γ on $\mathsf{relint}(\mathcal{R})$ if and only if there exists some convex $G : \mathcal{R} \to \mathbb{R}$ with subgradients $\{dG_r\}_{r \in \mathcal{R}}$, and some linear $\ell \in \mathsf{Lin}(\mathcal{V} \to \mathbb{R})$, such that for all $r \in \mathsf{relint}(\mathcal{R})$ and $t \in \mathcal{T}$,*

$$\mathfrak{A}(r)(t) \;=\; G(r) + dG_r(\Gamma(t) - r) + \ell(t). \tag{3.37}$$

Proof. It is trivial that the given form is truthful, as observed above: ℓ plays no part in the elicitation, and the remainder is truthful by Theorem 2.1. For the converse, given some $\mathfrak{A}$ which weakly elicits Γ on $\mathsf{relint}(\mathcal{R})$, we obtain an ℓ from Lemma 3.22 such that for all $r \in \mathsf{relint}(\mathcal{R})$, the score $\mathfrak{A}'(r)(t) \doteq \mathfrak{A}(r)(t) - \ell(t)$ depends on t only through Γ. To see this, fix $t_r \in \Gamma_r$ and let $\mathfrak{A}^{\mathcal{R}}(r')(r) \doteq \mathfrak{A}'(r')(t_r)$; then for $t \in \Gamma_r$,

$$\mathfrak{A}'(r')(t) - \mathfrak{A}^{\mathcal{R}}(r')(\Gamma(t)) = \mathfrak{A}(r')(t) - \ell(t) - \mathfrak{A}'(r')(t_r) + \ell(t_r) = 0,$$

where the last equality uses Lemma 3.22. We now note that as the choice of t_r above was arbitrary, we may take a linear right inverse of Γ, satisfying $t_{r+\alpha r'} = t_r + \alpha t_{r'}$ (this can be done by Lemma 3.21). Now $\mathfrak{A}^{\mathcal{R}}$ satisfies the conditions of Theorem 2.1 for type space $\mathsf{relint}(\mathcal{R})$, from which the form (3.37) follows. □

We conclude with a few remarks. First, note that the restriction to $\mathsf{relint}(\mathcal{R})$ is not merely for convenience, as the following example shows.

Example 3.3. *Let $\mathcal{T} = \{(t_1, t_2) \in [0,1]^2 : t_2 \leq t_2\}$, and $\Gamma(t) = t_1$. Then the following affine score elicits Γ:*

$$\mathfrak{A}(r)(t) = \begin{cases} 2rt_1 - r^2 & \text{if } r > 0 \\ -t_2 & \text{if } r = 0 \end{cases}. \tag{3.38}$$

To see this, note that $\mathfrak{A}(0)(t) \leq 0$ for all t, so when $t_1 > 0$ reporting $r = t_1$ is optimal. For $t_1 = 0$, the only type in $\mathcal{T}$ also has $t_2 = 0$, and hence $r = 0$ strictly dominates. Thus, we have strict elicitation. But $\mathfrak{A}$ cannot be written in the form (3.37) as $\mathfrak{A}$ depends on t_2 in a way that is not constant in r.

Intuitively, Theorem 3.23 is using the fact that Γ forces $\mathfrak{A}(\Gamma(t))(t)$ to be flat along level sets of Γ, which tells one a lot about the subgradients on interior points. However, Example 3.3 shows that all bets are off for boundary points. Put another way, if r is in the (relative) interior, then the affine score has to be "well-behaved" for reports nearby. But if r is up against the boundary of $\mathcal{R}$, then the score can do something more extreme, since it only has to be well-behaved on one side of r.

We briefly note also that Theorem 3.23 applies to some nonlinear properties as well. For any invertible $\psi : \mathcal{R} \to \mathcal{R}'$ and linear property Γ, an affine score $\mathfrak{A} : \mathcal{R}' \to \mathcal{A}$ weakly elicits $\Gamma'(t) \doteq \psi(\Gamma(t))$ if and only if

$$\mathfrak{A}(r')(t) \;=\; G(\psi^{-1}(r')) + dG_{\psi^{-1}(r')}(\Gamma(t) - \psi^{-1}(r')) + \ell(t), \tag{3.39}$$

of course under the same conditions as Theorem 3.23.

Finally, we remark that the proof above may have room for simplification.

Question 3.2. *For linear Γ, can we show $\mathfrak{A}(r)(t) = \mathfrak{A}^{\mathcal{R}}(r)(\Gamma(t)) + \ell(t)$ for some ℓ, without resorting to differentiation arguments?*

3.4.3 Nonlinear properties

We now turn to the general (i.e. nonlinear) functional property case, though our results will be far from general. We will need to make strong smoothness and geometrical assumptions. We restrict to the scoring rule setting with finitely many outcomes $\mathcal{O} = [n]$, so that $\mathcal{T} = \mathcal{P} \subseteq \Delta_n$ is a convex set of distributions. For compact notation, we write our scores as $\mathfrak{S} : \mathcal{R} \to \mathbb{R}^n$ rather than the traditional $\mathfrak{S} : \mathcal{R} \times \mathcal{O} \to \mathbb{R}$; thus, the expected score of report r under belief p is $\mathfrak{S}(r) \cdot p$.

Our result ties in nicely with the scoring rules literature, so we first review the relevant work. Recall from §1.2.1 the price rate argument of Savage [92]: to obtain a truthful report of an agent's value for a commodity, simply offer to sell one unit of the commodity at each of a sequence of prices, in even intervals between 0 and the reported price π. One easily sees that "one unit" could easily be replaced by "any positive amount" and the elicitation would still be valid. In fact, Savage points out that making this continuous and selecting some positive amount $\lambda(\pi)$ yields the representation

$$\mathfrak{S}(\pi', \pi) = \int_0^{\pi'} \lambda(\alpha)(\pi - \alpha) d\alpha, \tag{3.40}$$

since the agent gains value $\lambda(\alpha)\pi$ for each transaction but pays $\lambda(\alpha)\alpha$.

This integral representation of scoring rules has come up again and again; that of Schervish [93] is perhaps the most famous, but it also appears in Osband [78], Lambert et al. [65, 63], and Gneiting [50]. In all cases, it is noted that one may take some set of proper *base scores* (usually not strictly proper) and merely take a weighted integral over them to obtain another proper scoring rule.

The result of Lambert et al. [65] is particularly relevant here, as it addresses nonlinear scalar properties. They show that a scoring rule is proper for $\Gamma : \mathcal{P} \to \mathbb{R}$ if and only if one can write

$$\mathfrak{S}(r') = \int_{r_0}^{r'} \lambda(r) v(r) dr, \tag{3.41}$$

where $\lambda(r')$ is any nonnegative function and v is determined uniquely by Γ alone. The authors also extend this result to higher dimensional $\mathcal{R}$, but under very restrictive conditions. In particular, for a linear property Γ, their result does not cover the score $\mathfrak{S}(r) = G(r) + \nabla_r G \cdot (\Gamma(p) - r)$ where $G(r) = r^\top M r$ for non-diagonal p.s.d. $M \in \mathbb{R}^{k \times k}$; that is, when G is a quadratic form which is not axis-aligned.

As further motivation, consider the form we obtain when applying the linear characterization above (Theorem 3.23). We spell this out in §4.2.1, but briefly, we find that

$$\mathfrak{S}(r', p) = G(r) + \nabla_r G \cdot (\Gamma(p) - r) + \ell(p) \tag{3.42}$$

for some linear $\ell : \mathcal{P} \to \mathbb{R}$. Now assume that $\mathfrak{S}$ is differentiable and G is twice differentiable; then we have

$$\nabla_r \mathfrak{S}(\cdot, p) = \nabla_r G + (\Gamma(p) - r)^\top \nabla_r^2 G - \nabla_r G = (\Gamma(p) - r)^\top \nabla_r^2 G. \tag{3.43}$$

Hence, letting $\Lambda(r) = \nabla_r^2 G$ and $V(r) = (A - r\mathbb{1}^\top)$, where $\Gamma(p) = Ap$, we have

$$\mathfrak{S}(r') = \int_{r_0}^{r'} \Big(\Lambda(r)\ V(r)\Big)^\top dr. \tag{3.44}$$

Thus, we again have V determined uniquely by Γ and Λ an (essentially) arbitrary positive semidefinite matrix function (p.s.d. as G is convex), and we can write any $\mathfrak{S}$ as the integral over these choices.

We will show that any continuously differentiable proper scoring rule for a *nonlinear* property Γ can be represented in a similar way, provided it satisfies the condition below. The crucial insight is in the construction of V. We will see that just as $A = \nabla_p \Gamma$ for all $p \in \mathcal{P}$ in the linear case, using $\nabla\Gamma$ for V will also work in the nonlinear case, evaluated at a particular choice of representatives $\hat{p}(r)$.

$$\text{Condition:} \quad \begin{array}{l} \Gamma \text{ is elicitable, differentiable, and has convex maximal level} \\ \text{sets of dimension } \dim(\mathcal{P}) - \dim(\mathcal{R}) = n - k - 1. \end{array} \tag{3.45}$$

As usual, we set $\mathcal{R} \doteq \Gamma(\mathcal{P})$. We say a set S is *convex maximal* (in $\mathcal{P}$) if S is the intersection of an affine subspace and $\mathcal{P}$. In other words, S is maximally convex in $\mathcal{P}$, in the sense that it contains all points in its affine extension which lie in $\mathcal{P}$.

We first prove a useful lemma which lets us decompose the derivative of $\mathfrak{S}$.

Lemma 3.24. *Let continuous $V(r) \in \mathbb{R}^{k \times n}$ be given such that $\ker V(r) = \operatorname{span}(\Gamma^{-1}(r))$ for all $r \in \mathcal{R}$. Then for any continuously differentiable proper scoring rule $\mathfrak{S}$ for Γ, one can write $(\nabla_r \mathfrak{S})^\top = \Lambda(r)V(r)$ for some continuous $\Lambda(r) \in \mathbb{R}^{k \times k}$.*

Proof. By propriety, any $p \in \Gamma^{-1}(r)$ must satisfy $(\nabla_r \mathfrak{S})^\top p = \mathbf{0}$ (but not the converse need not hold). Hence,

$$\ker V(r) = \operatorname{span}(\Gamma^{-1}(r)) \subseteq \ker (\nabla_r \mathfrak{S})^\top.$$

But now we have

$$\operatorname{im} V(r)^\top = (\ker V(r))^\perp \supseteq (\ker (\nabla_r \mathfrak{S})^\top)^\perp = \operatorname{im} \nabla_r \mathfrak{S},$$

meaning each column of $\nabla_r \mathfrak{S}$ can be expressed as a linear combination of the rows of $V(r)$. Folding the coefficients of these linear combinations into $\Lambda(r)$ gives us the $(\nabla_r \mathfrak{S})^\top = \Lambda(r)V(r)$. For continuity of Λ, note that by assumption on Γ, we have $\dim(\Gamma_r) = n - k - 1$, and thus $\dim(\ker V) = n - k$. Hence, by classic results regarding the Moore-Penrose pseudoinverse [83, Thm 4.2], the pseudoinverse $V(r)^+$ of V is continuous, so $\Lambda(r) = \nabla_r \mathfrak{S}^\top V(r)^+$ is continuous as the product of continuous functions. □

Theorem 3.25. *Let $\Gamma : \mathcal{P} \to \mathcal{R} \subseteq \mathbb{R}^k$ satisfy condition (3.45) and let $\mathfrak{S} : \mathcal{R} \to \mathbb{R}^n$ be a continuously differentiable scoring rule which is proper for Γ. Then $\mathfrak{S}$ satisfies*

$$(\nabla_r \mathfrak{S})^\top = \Lambda(r)V(r) \tag{3.46}$$

for $V(r) \in \mathbb{R}^{n \times k}$ determined by Γ alone, and $\Lambda(r) \in \mathbb{R}^{k \times k}$ positive semi-definite.

Proof. Let $\hat{p} : \mathcal{R} \to \mathcal{P}$ be any continuously differentiable function such that $\Gamma \circ \hat{p} = \mathrm{id}_{\mathcal{R}}$, which exists by the inverse function theorem, and define

$$V(r) = \left(\nabla_{\hat{p}(r)} \Gamma\right) \left(I - \hat{p}(r) 1^\top\right). \tag{3.47}$$

Note that this choice of $V(r)$ depends only on Γ (and our choice of $\hat{p}$, which in turn depends only on Γ).

By construction,

$$q \in \ker V(r) \iff \left(\nabla_{\hat{p}(r)}\Gamma\right)(q - \hat{p}(r)) = \mathbf{0}.$$

Since the level sets of Γ are convex maximal, we see that the derivative of $\nabla_{\hat{p}(r)}\Gamma$ in the direction $(q-\hat{p}(r))$ is zero precisely when $\Gamma(q) = \Gamma(\hat{p}(r)) = r$. Thus, $\ker V(r) \subseteq \text{span}(\Gamma^{-1}(r))$. Now by Lemma 3.24, we have some continuous $\Lambda : \mathcal{R} \to \mathbb{R}^{k\times k}$ such that $(\nabla_r \mathfrak{S})^\top = \Lambda(r)V(r)$, and it remains to show that $\Lambda(r)$ is positive semi-definite.

For a contradiction, fix $r \in \mathcal{R}$ and any unit-length $v \in \mathbb{R}^k$ such that $v^\top \Lambda(r) v < 0$. By continuity of $\Lambda(\cdot)$, there exists some $\epsilon > 0$ such that

$$\forall\, r' \in B_\epsilon(r) \;\; \forall\, v' \in B_\epsilon(v) : \quad v^\top \Lambda(r') v' < 0, \tag{3.48}$$

where $B_\epsilon(\cdot)$ denotes the L_2 ball of size ϵ. Now define

$$f(\alpha) = \mathfrak{S}(r + \alpha v) \cdot \hat{p}(r).$$

By the above, the derivative of f is given by

$$\begin{aligned} df(\alpha) &= v^\top (\nabla_r \mathfrak{S})^\top \hat{p}(r) \\ &= v^\top \Lambda(r+\alpha v) V(r+\alpha v) \hat{p}(r) \\ &= v^\top \Lambda(r+\alpha v) \left(\nabla_{\hat{p}(r+\alpha v)}\Gamma\right)(\hat{p}(r) - \hat{p}(r+\alpha v)). \end{aligned}$$

By properties of directional derivatives and the chain rule,

$$\begin{aligned} &\lim_{\alpha\to 0} \left(\nabla_{\hat{p}(r+\alpha v)}\Gamma\right)\left(\frac{\hat{p}(r) - \hat{p}(r+\alpha v)}{\alpha}\right) \\ &= \left(\nabla_{\hat{p}(r)}\Gamma\right)\left(\lim_{\alpha\to 0} \frac{\hat{p}(r) - \hat{p}(r+\alpha v)}{\alpha}\right) \\ &= \left(\nabla_{\hat{p}(r)}\Gamma\right)(\nabla_r \hat{p})\,(-v) \\ &= \nabla_r(\Gamma \circ \hat{p})(-v) \\ &= -v. \end{aligned}$$

Hence, there exists some $\delta > 0$, $\delta < \epsilon$, such that for all $0 < \alpha \le \delta$ we have

$$v_\alpha := -\alpha^{-1}\left(\nabla_{\hat{p}(r+\alpha v)}\Gamma\right)(\hat{p}(r) - \hat{p}(r+\alpha v)) \in B_\epsilon(v).$$

We also have $r + \alpha v \in B_\epsilon(r)$ since $\alpha < \epsilon$, so by (3.48),

$$df(\alpha) = v^\top \Lambda(r+\alpha v)(-\alpha v_\alpha) > 0.$$

Thus, f is increasing for $0 < \alpha \le \delta$, so we have

$$\mathfrak{S}(r) \cdot \hat{p}(r) = f(0) < f(\delta) = \mathfrak{S}(r + \delta v) \cdot \hat{p}(r),$$

which contradicts $\mathfrak{S}$ being proper for Γ. □

We now see that, in fact, the integral representation (3.41) above also holds for general (well-behaved) vector-valued properties as well. For any smooth $\mathfrak{S}$ eliciting Γ satisfying (3.45), we may write

$$\mathfrak{S}(r') = \int_{r_0}^{r'} \Big(\Lambda(r)\ V(r)\Big)^\top dr; \tag{3.49}$$

this is exactly the same formula (3.44) that we motivated from the linear case! Intuitively, one should be able to find a path $r(t)$ from r_0 to r' such that $t \mapsto r(t)^\top \Lambda(r(t))V(r(t))p$ is monotonically decreasing, just as with Savage's price rate formula (3.40), with the zero being precisely when $r(t) = \Gamma(p)$. In this sense, we are essentially generalizing the "offer to sell units of a commodity at increasing prices" trick to vector-valued nonlinear settings.

Note that we have not said anything at all about which properties Γ are elicitable; this is a major open question. By the above intuition, we might expect that the elicitable properties are precisely those which admit a path $r(t)$ yielding such monotone behavior. An important first step to answering this question is understanding when the form $V(r_0)$ for fixed r_0 can itself be viewed as a (non-strictly) proper scoring rule for Γ, i.e., for what Γ is $V(r_0)$ a base score?

Also, we have not characterized the functions Λ which the designer may choose from to construct $\mathfrak{S}$. This is in essence the analog of the "degrees of freedom" discussion from §3.4.1 for finite properties. We conjecture that Λ is the Hessian of a twice continuously differentiable convex function $G : \mathcal{R} \to \mathbb{R}$, just as in the linear case. If this were true, then scoring rules would in some sense be "nonlinearly transformed" Bregman divergences. This intuition seems consistent with the examples given by Gneiting [50].

Finally, it is interesting to find sufficient conditions for the level sets of Γ to be convex maximal. One such condition is that Γ itself be monotone, in the sense that on any line $L : \mathbb{R} \to \mathcal{P}$ and for any direction $v \in \mathbb{R}^k$, the function $(\Gamma \circ L) \cdot v = (\alpha \mapsto \Gamma(L(\alpha)) \cdot v)$ is monotone.

Lemma 3.26. *If Γ is monotone in the sense above, the level sets Γ_r are convex maximal in $\mathcal{R}$.*

Proof. By monotonicity of Γ, the functions $\gamma^v(p) \doteq \Gamma(p) \cdot v$ have convex level sets (take p, p' such that $\Gamma(p) \cdot v = \Gamma(p') \cdot v$; the function $\alpha \mapsto \Gamma(\alpha p + (1-\alpha)p') \cdot v$ is monotone). Now by an argument of Lambert [63] given in Step 2 of the proof of Theorem 5, we have that the level sets of γ^v are convex maximal; we briefly sketch the argument here. The goal is to show that $\{p : \gamma^v(p) < x\}$ and $\{p : \gamma^v(p) > x\}$ are convex, and then invoke classic results in geometry to conclude that $\gamma^v_x \doteq \{p : \gamma^v(p) = x\}$ is a hyperplane intersected with $\mathcal{P}$. Letting $p, p' \in \mathcal{P}$ with $\gamma^v(p) > \gamma^v(p') > x$, let $p_\alpha = \alpha p + (1-\alpha)p'$ and define $f(\alpha) = \gamma^v(p_\alpha)$. By continuity, we have $[\gamma^v(p), \gamma^v(p')] \subseteq f([0,1])$, and by convexity of the level sets of γ^v we conclude $[\gamma^v(p), \gamma^v(p')] = f([0,1])$.

Now we have that γ^v_x is convex maximal for all v and x. To conclude, we note that

$$\Gamma_r = \bigcap_{v \in \mathbb{R}^k} \gamma^v_{r \cdot v}, \tag{3.50}$$

and thus Γ_r must be convex maximal as well. □

3.5 Future work

Many questions in the literature on properties remain open. Most notable is the characterization of elicitable nonlinear and multidimensional properties and their scores — real-valued distributional properties are covered in [65, 63] and the linear vector-valued case is treated above. We hope that the preliminary results and intuition from § 3.4.3 will yield a useful characterization in this case.

Another interesting direction is for non-functional properties: aside from the finite $\mathcal{R}$ case, the vast majority of the literature to our knowledge assumes that Γ is a function (specifying a single correct report for each type). The generality of Theorem 3.2 may prove useful in exploring non-functional settings as well. A result requiring few regularity conditions on Γ would be useful in domains such as statistics where natural properties like the median cannot in general be expressed as functions.

Chapter 4

Eliciting means of distributions

4.1 Introduction

In this chapter we draw strong connections between four seemingly distant concepts: scoring rules, exponential families, prediction markets, and Bregman divergences. Many pairs of these concepts have been examined before, but by fitting them all under one roof, we are able to make a much clearer picture of how these objects are interrelated. We first outline our findings from a high level, then discuss previous work and give the formal background.

4.1.1 Overview

The goal of this chapter is to draw connections between scoring rules, (generalized) exponential families, prediction markets, and Bregman divergences. While we will give formal definitions below, for now the reader may think of exponential families as maximum entropy distributions under a mean constraint of a particular statistic. (For the other concepts see §1.2.1, §1.2.2, and §3.3.3.)

Our exploration into these connections is motivated by the role of convexity in each — the Bayes risk of scoring rules, the cost function of a prediction market, the cumulant of exponential families, and the generating function of a Bregman divergence — and the mysterious appearance of a map which we call ϕ whose mean under some fixed distribution is central to each theory. Beyond these notions, however, we are motivated by several natural questions:

- When can a scoring rule for the mean of a random variable be written in terms of a scoring rule for the entire underlying distribution?
- When can an incomplete prediction market be run as a special case of a complete one?
- What is the (generalized) relative entropy of two members of a (generalized) exponential family, in terms of their parameters?

- Are prediction markets more or less expressive than scoring rules?
- How is the statistic of an exponential family related to the random variable in a proper scoring rule for an expectation? To the payoff function of an incomplete prediction market?
- What is the net payoff of a trade in a prediction market, expressed in terms of the prices before and after the trade?

We will answer all of these questions in this chapter, as well as several questions that one might not think to ask, such as the relationship between incomplete prediction markets and exponential families.

In short, the crux of our conclusions is that all four objects are in some sense equivalent. Figure 4.1 gives this high-level depiction of our results; see Table 4.1 for a more detailed retrospective view of the chapter. Each "$\Longleftrightarrow$" in Figure 4.1 represents a bijection between the two constructs, that is, a mapping from one to the other and back. Moreover, under certain global assumptions discussed below, the diagram commutes, in the sense that composing bijections on edges yields the remaining bijections. In all cases, the bijections go beyond merely "counting", but say something deeper about the relationship between the concepts, enabling us to answer the above questions.

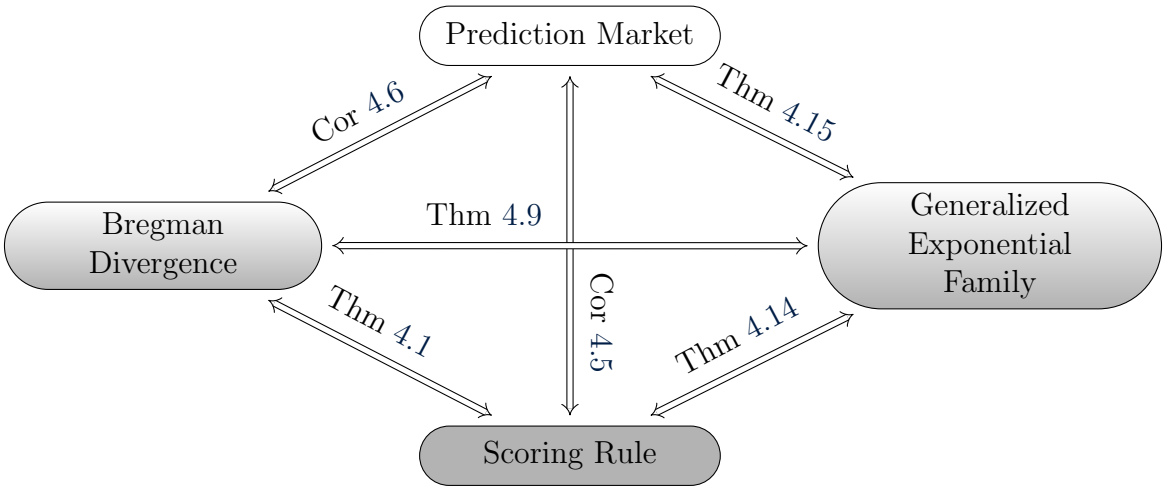


Figure 4.1: A four-way equivalence. Gray denotes primal objects (distributions), while clear denotes dual objects (random variables) — divergences and GEFs can act as both.

4.1.2 Previous work

Several connections between the four concepts we consider have been discovered before. To start, the relationship between Bregman divergences and scoring rules for expected values

dates all the way back to Savage's 1971 paper, when he introduced what is now called the "Savage representation" (1.2), which is essentially a Bregman divergence [92].[1] Since then his representation has been further formalized and extended, by e.g. Gneiting and Raftery [51] and Gneiting [50].

Even outside the scoring rules literature, it has been observed in machine learning that the minimizer of the empirical average of Bregman divergences is the mean of its argument (a generalization of the *principle of least squares*), and Banerjee et al. [15] give conditions under which Bregman divergences are the *only* loss function with this property. While using a somewhat different terminology, one can also find a very thorough treatment in Reid et al. [84] and Vernet et al. [96].

Hence, it is fair to say that our result that scoring rules are in some sense equivalent to (generalized) Bregman divergences is far from groundbreaking. However, there is value in the generality of our result; to our knowledge, it is the most general of any in the literature, allowing for quite general vector spaces and very few regulatory conditions. In particular, nearly all previous results for the expected value case (i.e. eliciting $\Gamma(p) = \mathbb{E}_p[\phi]$ for some $\phi : \mathcal{O} \to \mathcal{R}$) require some sort of differentiablity. Some authors, e.g. [15, 50], also make the assumption that one observes only the value of the random variable, not the randomness generating it (i.e. $\phi(o)$ rather than o itself, or equivalently $\mathcal{O} \subseteq \mathcal{R}$ and $\phi = \mathrm{id}_{\mathcal{R}}$), thus making the characterization less general. Gneiting [50] points out that this previous work all involve smoothness conditions, and states, "A challenging, nontrivial problem is to unify and strengthen these results, both in univariate and multivariate settings" — in the present chapter, we aim to do just that.

Adding prediction markets to the story, we already saw in §1.2.2 Hanson's equivalence between the scoring rule and prediction market form of the LMSR [56]. More recently, the {scoring rule, prediction market, Bregman divergence} trio has been discussed in Abernethy et al. [3]. There the authors point out that the net payoff of a trade in a prediction market can be written as a Bregman divergence between the final outcome and the prices before and after the trade. They also show a general equivalence theorem between market scoring rules and prediction markets for the complete market setting, showing that each model can be simulated with the other. Our results will generalize these, in removing differentiability assumptions and extending the equivalence to the incomplete market setting.

Exponential families have a rich history and literature, dating back to the 1930s. We give a very brief overview of their general theory and derivation in §4.3. Many properties of exponential families are known, but for our exploration, the most relevant is the connection to Bregman divergences. It has been observed in particular that the relative entropy between two members of an exponential family is the Bregman divergence with respect to the cumulant of the two corresponding parameters (cf. Amari [6], Azoury & Warmuth [12], and Nielsen & Nock [76]). Banerjee et al. [16] leverage this identity to show a bijection between Bregman divergences and exponential families.

[1] In fact, the L function he considers, which we called a *score divergence* in §3.3.3, is precisely a Bregman divergence in his setting.

Another key property of exponential families is that they are *maximum entropy* distributions; given a statistic, the distribution with a particular mean of the statistic which is of maximum entropy forms an exponential family [18]. In 2004, Grünwald and Dawid [54] introduced *generalized* exponential families (GEFs) as maximum entropy distributions for other entropy functions — that is, they noted that one may *define* classical exponential families as maximum entropy distributions for the usual Shannon entropy, thus making the generalization to other notions of "entropy" immediate. We will work with these GEFs, though our definition will slightly depart from theirs by mirroring the exponential families literature more closely, thereby illuminating connections to that literature more easily. Some aspects of our study of GEFs are inspired by unpublished work of Sébastien Lahaie.

4.1.3 Definitions and notation

We now give the formal definitions that will be used throughout the chapter. Note however that we defer the formal introduction of generalized exponential families to §4.3. Throughout, $\mathcal{O}$ is a (possibly infinite) set of outcomes, $\mathcal{P} \subseteq \Delta(\mathcal{O})$ is a convex set of probability measures, and $\mathcal{R}$ is a convex report space. Every result will involve a function $\phi : \mathcal{O} \to \mathcal{R}$, though this function plays many roles: a statistic, a payoff function, a random variable, and a link function.

For the majority of the chapter, owing to our heavy use of convex conjugate duality, we will work with a *dual pair* of vector spaces, and indeed a pair of dual pairs in § 4.3; see Definition 3.6 and [5, §5.14]. The dual pair $(\mathcal{V}, \mathcal{V}^*)$ captures the report space $\mathcal{R} \subseteq \mathcal{V}$. The dual pair $(\mathcal{W}, \mathcal{W}^*)$ is between probability distributions $\mathcal{P} \subseteq \mathcal{W}$ and random variables $\mathcal{W}^*$, with the standard bilinear form $\langle p, f \rangle \doteq \mathbb{E}_p[f]$. Note that when we restrict to the dual pair setting, we must have $\mathbb{E}_p[f] \in \mathbb{R}$ for all $p \in \mathcal{W}$ and $f \in \mathcal{W}^*$. This level of abstraction allows us to work with measure spaces, such as the pair $\langle C(X), ca(X) \rangle$ where X is a compact subset of $\mathbb{R}^n$.[2] That said, a reader uninterested in applications to measure spaces may safely assume that a base measure ν on $\mathcal{O}$ has been chosen, and identify $p \in \mathcal{P}$ with its density function.

Each result in this chapter has its own assumptions, on the various functions and spaces involved, which is also summarized at the end in Table 4.1. However, all results in this chapter hold under the assumption that we are working in a dual pair of vector spaces and the convex functions are all proper, strictly convex, and differentiable (though the discussion gives a more precise account of these "global" necessary conditions). A reader more interested in the conceptual rather than technical content may safely make this single set of assumptions throughout, perhaps even further assuming $\mathcal{V} = \mathcal{V}^* = \mathbb{R}^k$ for concreteness.

We now introduce the characters of our story, starting with divergences.

Definition 4.1. *A* generalized Bregman divergence *on space X is a function $D_{G,dG} : X \times X \to \overline{\mathbb{R}}$ given by*

$$D_{G,dG}(x, x') = G(x) - G(x') - dG_{x'}(x - x'), \tag{4.1}$$

[2]Here $C(X)$ denotes continuous functions from X to $\mathbb{R}$ and $ca(X)$ denotes the space of bounded and countably additive signed measures on the sigma algebra associated with X (here the Borel algebra).

where $G : \mathsf{conv}(X) \to \overline{\mathbb{R}}$ *is convex with* $G(X) \subseteq \mathbb{R}$*, and* dG *is a subgradient of* G*. If* G *is differentiable on* X *we simply write* D_G *and call* D_G *a* Bregman divergence*.*

Note that, as mentioned in §3.3.3, when G is continuously differentiable, the form (4.1) is simply called a *Bregman divergence*. Hence, Definition 4.1 is merely a natural extension to the nondifferentiable case, and has been studied in machine learning (cf. [60]). In this chapter, our spaces will be convex, so we will always have $G : X \to \mathbb{R}$ in the above.

The following are slightly reformulated definitions of scoring rules and prediction markets, so that we may more easily use the results of Chapter 2 and Chapter 3 in our exploration. Note that in both cases the corresponding affine score is given by the expected payoff under the "type" $p \in \mathcal{P}$, e.g. $\mathfrak{A}(r)(p) = \mathfrak{S}(r, p)$. Throughout the chapter we will refer to both $\mathfrak{S}$ and $\mathfrak{P}$ as affine scores, having this relationship in mind.

Definition 4.2. *Given outcome space* $\mathcal{O}$*, report space* $\mathcal{R}$*, and set of probability measures* $\mathcal{P} \subseteq \Delta(\mathcal{O})$*, a* scoring rule *is a function* $\mathfrak{S} : \mathcal{R} \times \mathcal{O} \to \overline{\mathbb{R}}$*. We write* $\mathfrak{S}(r, p) \doteq \mathbb{E}_{o\sim p}[\mathfrak{S}(r, o)]$ *to denote the expected score. We say* $\mathfrak{S}$ (weakly) elicits *a property* $\Gamma : \mathcal{P} \to \mathcal{R}$ *if for all* $p \in \mathcal{P}$ *and all* $r \in \mathcal{R}$*,*

$$\mathfrak{S}(r, p) \leq \mathfrak{S}(\Gamma(p), p). \tag{4.2}$$

Definition 4.3. *Given outcome space* $\mathcal{O}$*, price space* $\mathcal{R}$*, share space* $\mathcal{Q}$*, payoff function* $\phi : \mathcal{O} \to \mathcal{R}$*, and set of probability measures* $\mathcal{P} \subseteq \Delta(\mathcal{O})$*, a* prediction market *is a function* $\mathfrak{P} : \mathcal{Q} \times \mathcal{O} \to \overline{\mathbb{R}}$ *defined by*

$$\mathfrak{P}(q, o) \doteq \langle q, \phi(o)\rangle - C(q), \tag{4.3}$$

where $C = G^*$ *for some convex* $G : \mathsf{conv}(\mathcal{R}) \to \overline{\mathbb{R}}$ *with* $G(\mathcal{R}) \subseteq \mathbb{R}$ *and* $\partial G(\mathcal{R}) = \mathcal{Q}$*. As with scoring rules, we write* $\mathfrak{P}(q, p) \doteq \mathbb{E}_{o\sim p}[\mathfrak{P}(q, o)]$ *to denote the expected payout.*

A reader who is even remotely familiar with prediction markets may find Definition 4.3 somewhat confusing, as it is lacking the sequential nature of a prediction market mechanism — note that the goal here is not to capture the dynamic mechanism of buying and selling, but to capture the "one-round" incentives as an affine score. The standard framework discussed in § 1.2.2 can be recovered by considering a trade $r \in \mathbb{R}^k$ when the market is at state $q \in \mathbb{R}^k$. The cost of this purchase is $C(q + r) - C(q)$, and upon outcome o being revealed, the trade pays off $\phi(o) \cdot r$. The net payoff of this trade then is just $\mathfrak{P}(q + r, o) - \mathfrak{P}(q, o)$. We can thus capture the usual dynamic prediction market mechanism by taking differences of the $\mathfrak{P}$ function: for a trade sequence yielding market state vectors $q_0, q_1, \ldots, q_N$, the net payoff of the agent responsible for trade $q_i \mapsto q_{i+1}$, upon outcome o being revealed, is $\mathfrak{P}(q_{i+1}, o) - \mathfrak{P}(q_i, o)$.

As stated, (affine scores for) prediction markets are much more restricted in their dependence on o than scoring rules are. In particular, a prediction market can only depend on o through ϕ. To "even the playing field," we introduce the notion of a *fair score*, which in essence imposes this condition on scoring rules as well.

Definition 4.4. *A scoring rule* $\mathfrak{S} : \mathcal{R} \times \mathcal{O} \to \overline{\mathbb{R}}$ *is* Γ-fair *for property* Γ *if for all* $p, p' \in \mathcal{P}$ *such that* $\Gamma(p) = \Gamma(p') = r$*,* $\mathfrak{S}(r, p) = \mathfrak{S}(r, p')$*.*

4.2 Prediction market duality

In this section we consider the relationship between prediction markets and scoring rules, when viewed as affine scores. Are the scores $\mathfrak{S}$ and $\mathfrak{P}$ related in any way? Is one more expressive than the other, all else being equal? Along the way we will relate both scoring rules and prediction markets to divergences.

Of course we must formalize what it means for one affine score to be more or less expressive than another; for this we introduce the notion of a reduction.

Definition 4.5. *Affine score* $\mathfrak{A}_1 : \mathcal{R}_1 \to \mathcal{A}$ reduces to $\mathfrak{A}_2 : \mathcal{R}_2 \to \mathcal{A}$ *if there exists some map* $\varphi : \mathcal{R}_1 \to \mathcal{R}_2$ *such that for all* $r \in \mathcal{R}_1$ *and* $t \in \mathcal{T}$*,*

$$\mathfrak{A}_1(r)(t) = \mathfrak{A}_2(\varphi(r))(t).$$

If furthermore $\mathfrak{A}_2$ *reduces to* $\mathfrak{A}_1$*, we say the affine scores are* mutually reducible*.*

It will be useful to say two scoring rules are equivalent if the only difference is some predetermined payoff associated with each outcome, regardless of the agent's report. Intuitively, these rewards should have no bearing on the behavior of the agent, being entirely outside her control. Interestingly, this notion of equivalence was introduced as early as McCarthy [71], and is now known as *strong equivalence* in the scoring rules literature; see e.g. Gneiting and Raftery [51] and Dawid [39].

Definition 4.6. *Scoring rules* $\mathfrak{S} : \mathcal{R} \times \mathcal{O} \to \overline{\mathbb{R}}$ *and* $\mathfrak{S}' : \mathcal{R} \times \mathcal{O} \to \overline{\mathbb{R}}$ *are* equivalent *if for all* $o \in \mathcal{O}$*,* $\mathfrak{S}'(r, o) - \mathfrak{S}(r, o)$ *is a constant independent of* r*. In this case, we write* $\mathfrak{S}' \cong \mathfrak{S}$*.*

4.2.1 Scoring rules and divergences

As a warm-up, we show a strong relationship between scoring rules and divergences. This result is an easy application of Theorem 3.23, which gave a characterization of affine scores which weakly elicit linear properties. We simply apply it to the restriction that $\mathfrak{A}$ be a scoring rule, and note that divergences are trivially in one-to-one correspondence with the generating convex function (and choice of subgradients). Note that $\mathcal{P}$ here can be any convex set of probability measures, provided that $\mathbb{E}_p[\phi]$ is defined for all $p \in \mathcal{P}$ (see below). We implicitly assume regularity here; see Chapter 3 for details.

By standard properties of (generalized) Bregman divergences, such as nonnegativity, it is easy to see that the scoring rule defined by

$$\mathfrak{S}(r, o) = G(\phi(o)) - D_{G,dG}(\phi(o), r) + \ell(o) \tag{4.4}$$

elicits $\Gamma : p \mapsto \mathbb{E}_p[\phi]$; the expectation passes inside to yield $\mathfrak{S}(r, p) = f(p) - D_{G,dG}(\Gamma(p), r)$ for some f, and hence the optimal report is $r = \Gamma(p)$. Amazingly, this is in some sense the *only* form which elicits Γ. We can show this immediately, by applying Theorem 3.23 under the constraint that $\mathfrak{S}$ be a scoring rule.

Theorem 4.1. *Let $\mathcal{P} \subseteq \Delta(\mathcal{O})$ be convex, let map $\phi : \mathcal{O} \to \mathcal{V}$ be given, and set $\Gamma(p) = \mathbb{E}_p[\phi]$ and $\mathcal{R} = \Gamma(\mathcal{P})$. Then $\mathfrak{S}$ weakly elicits Γ on $\Gamma^{-1}(\mathsf{relint}(\mathcal{R}))$ if and only if there exists some convex $G : \mathcal{R} \to \mathbb{R}$ with subgradients $\{dG_r\}_{r \in \mathcal{R}}$, and some map $\ell : \mathcal{O} \to \mathbb{R}$, such that for all $r \in \mathsf{relint}(\mathcal{R})$ and $o \in \mathcal{O}$,*

$$\mathfrak{S}(r, o) = G(r) + dG_r(\phi(o) - r) + \ell(o). \tag{4.5}$$

Furthermore, from the form (4.5) we can easily construct a divergence, namely

$$D_{G,dG}(r, r') = \mathfrak{S}(r, p_r) - \mathfrak{S}(r', p_r), \tag{4.6}$$

where $p_r \in \Gamma_r$ is any distribution with $\Gamma(p_r) = r$. Note that this works even though $\ell \neq 0$ in general, since the ℓ terms cancel out. Hence, we have a strong equivalence between generalized Bregman divergences and scoring rules for linear properties. Note that the restriction to $\mathsf{relint}(\mathcal{R})$ is necessary, in the sense that $\mathfrak{S}$ need not be of the form (4.5) on the boundary of $\mathcal{R}$; see Example 3.3.

4.2.2 Prediction markets and scoring rules

We now turn to the relationship between prediction markets and scoring rules, which will in turn transfer to divergences by the above. We first show that, under very broad assumptions, scoring rules can be expressed as prediction markets.

Proposition 4.2. *Let $\mathcal{P} \subseteq \Delta(\mathcal{O})$ be an arbitrary set of probability measures, and $\mathcal{R} \subseteq \mathcal{V}$ be given, both spaces being potentially non-convex. Let map $\phi : \mathcal{O} \to \mathcal{V}$ and $G : \mathcal{R} \to \overline{\mathbb{R}}$ be convex with $G(\mathcal{R}) \subseteq \mathbb{R}$ be given. Then affine score $\mathfrak{S}(r, o) \doteq G(r) + dG_r(\phi(o) - r)$ reduces to $\mathfrak{P}(q, o) \doteq q(\phi(o)) - G^*(q)$.*

Proof. Take $\varphi(r) = dG_r$. Then by Lemma 3.11, we have $G^*(dG_r) = dG_r(r) - G(r)$. Hence,

$$\begin{aligned} \mathfrak{S}(r, o) &= G(r) + dG_r(\phi(o) - r) \\ &= \varphi(r)(\phi(o)) - G^*(\varphi(r)) \\ &= \mathfrak{P}(\varphi(r), o). \end{aligned} \tag{4.7}$$

□

In light of Theorem 4.1, we know that every scoring rule which elicits $\Gamma : p \mapsto \mathbb{E}_p[\phi]$ is equivalent to a rule of the form (4.7) on $\mathsf{relint}(\mathcal{R})$. Thus, modulo these details, we conclude that prediction markets are *at least as expressive* as scoring rules.

Proposition 4.2 relies only on very basic tools from convex analysis, and in essence follows directly from the definition of the conjugate. As mentioned at the end of §3.3.2, this reduction is just report duality, in the same way that any mechanism can be considered in "menu" form.

To show the converse, that prediction markets reduce to scoring rules, will require more care, and in general the conditions will be more stringent. Intuitively, this structure is needed

because many share vectors represent the same market maker price, so while it is easy to map share vectors to prices (and indeed this is in some sense baked into the prediction market framework), going from prices to share vectors in a consistent way is more challenging. Among other conditions, will we need to work in our paired spaces $(\mathcal{V}, \mathcal{V}^*)$.

In an effort to make the proof more modular, we introduce now the major condition required for this result to hold. Assume G and subgradient dG are fixed, and $C = G^*$.

$$\forall q \in \mathcal{Q},\ \exists r \in \partial C_q,\ \forall o \in \mathcal{O}, \quad \langle q - dG_r, r - \phi(o) \rangle = 0. \tag{4.8}$$

This condition (4.8) essentially says that mapping from share vectors to prices and back has no effect on net payoffs. Suppose a trader has belief $p \in \Delta(\mathcal{O})$ and buys vector q so that the market price becomes $r = \mathbb{E}_p[\phi]$; then for any realized outcome o the difference between the trader's expected payoff $\langle q, r \rangle$ and actual payoff $\langle q, \phi(o) \rangle$ is the same if we replace q by the "equivalent" share vector dG_r.

Theorem 4.3. *Given dual pair $(\mathcal{V}, \mathcal{V}^*)$ with $\mathcal{R} \subseteq \mathcal{V}$, let $G : \mathcal{R} \to \mathbb{R}$ be convex, proper, and l.s.c., and let $C = G^*$. If condition (4.8) holds, then the affine scores $\mathfrak{P}(q, o) \doteq \langle q, \phi(o) \rangle - C(q)$ and $\mathfrak{S}(r, o) \doteq G(r) + \langle \phi(o) - r, dG_r \rangle$ are mutually reducible.*

Proof. One direction has been proved in more generality in Proposition 4.2. For the converse, by the Fenchel–Moreau Theorem, given as Theorem 3.13, we immediately have $G^{**} = C^* = G$. Now for all $q \in \mathcal{Q}$ we define $\varphi(q) = r \in \partial C_q$ where r is the report guaranteed by condition (4.8). Applying Lemma 3.11 for C, we have $C(q) = \langle q, \varphi(q) \rangle - C^*(\varphi(q))$, whence

$$\begin{aligned} \mathfrak{P}(q, o) &= \langle q, \phi(o) \rangle - C(q) \\ &= C^*(\varphi(q)) + \langle q, \phi(o) - \varphi(q) \rangle \\ &= G(\varphi(q)) + \left\langle dG_{\varphi(q)}, \phi(o) - \varphi(q) \right\rangle \\ &= \mathfrak{S}(\varphi(q), o), \end{aligned}$$

where in the third equality we used $C^* = G$ and (4.8). □

Of course, it remains to show that condition (4.8) can be satisfied. We now give sufficient conditions, but it may be that (4.8) holds in other settings as well.

Lemma 4.4. *If G is proper and l.s.c., and $G(r; r' - r) = -G(r; r - r')$ for all $r \in \mathsf{relint}(\mathcal{R})$, $r \in \mathcal{R}$, then (4.8) holds for $r \in \mathsf{relint}(\mathcal{R})$ and q s.t. $\partial C_q \cap \mathsf{relint}(\mathcal{R}) \neq \emptyset$.*[3]

Proof. Fix q and $r \in \mathsf{relint}(\mathcal{R}) \cap \partial C_q \neq \emptyset$, and recall the definition of the convex conjugate for C:

$$C(q) = \sup_{r \in \mathcal{R}} \{\langle q, r \rangle - G(r)\}. \tag{4.9}$$

By Corollary 3.14 and Lemma 3.11 for C, we know that r must be a maximizer of the objective in (4.9). Now let $f(r) \doteq G(r) - \langle q, r \rangle$ denote the (negative) of this objective, and

[3]Recall that $g(x; d)$ is the directional derivative of g at x in direction d.

note that f is convex. Since r maximizes f over all $r \in \mathcal{R}$, we must in particular have $f(r; v) = 0$ for all $v \in \mathcal{R} - \{r\}$; otherwise, by assumption either $f(r; v)$ or $f(r; -v)$ would be strictly positive, contradicting optimality of r.

A standard result of convex analysis is that $g(x; v) = \sup\{\langle v, d\rangle : d \in \partial g_x\}$ for proper convex g and $x \in \mathsf{relint}(\mathsf{dom}(G))$ [89, Thm 23.4]. By our assumption on G however, note that $G(r; v) = -G(r; -v)$ implies that $\{\langle v, d\rangle : d \in \partial G_r\}$ is a singleton for each r. The same logic applies to f, since $\partial f_r = \{q\} - \partial G_r$. Thus, we have for all $d \in \partial G_r$ and all $r' \in \mathcal{R}$,

$$0 = f(r; r' - r) = \langle r' - r, q - d\rangle \,, \tag{4.10}$$

completing the proof. □

Putting the sufficient conditions from Lemma 4.4 together with Theorem 4.3, we now have a class of prediction markets and scoring rules which are mutually reducible.

Corollary 4.5. *Let $G : \mathcal{R} \to \mathbb{R}$ be convex, and let $C = G^*$. If additionally G is proper and l.s.c. and and $G(r; v) = -G(r; -v)$ for all $r \in \mathsf{relint}(\mathcal{R})$, $v \in \{r\} - \mathcal{R}$, then the affine scores $\mathfrak{P}(q, o) \doteq \langle q, \phi(o)\rangle - C(q)$ and $\mathfrak{S}(r, o) \doteq G(r) + dG_r(\phi(o) - r)$ are mutually reducible on $\mathcal{Q}' = \{q \in \mathcal{Q} : \partial C_q \cap \mathsf{relint}(\mathcal{R}) \neq \emptyset\}$ and $\mathcal{R}' = \mathsf{relint}(\mathcal{R})$.*

Once again applying Theorem 4.1, we now see that any *fair* scoring rule which elicits a linear property Γ is mutually reducible with a prediction market on the relative interior of the report space. Stepping back from our world of affine scores, Corollary 4.5 gives broad conditions under which the *market scoring rules* discussed in §1.2.2 have exactly the same expressiveness as incomplete prediction markets, thereby generalizing the equivalence result of Abernethy et al. [3, Thm 8.2] to the incomplete setting. Hence, we have come full circle: Hanson introduced the market maker prediction market as a scoring rule [56], which was "dualized" to a share-based cost function framework by Chen and Pennock [82], and further extended to incomplete setting by Abernethy et al. [1, 3]; we now have seen that the incomplete share-based market makers are again the same as market scoring rules for a linear property.

It is worth noting that the directional differentiability condition of G is a mild one from the perspective of prediction market design, as it essentially implies that share vectors q have "unique instantaneous prices." This is often a desirable property in and of itself, since otherwise prices will appear to jump as agents trade, or the "market consensus" price could be a range of values instead of a single point.

We now take one step further to give a direct connection between prediction markets and divergences.

Corollary 4.6. *Under the same conditions as Corollary 4.5, there exists an invertible $\varphi : \mathcal{R}' \to \mathcal{Q}'$ such that we may write*

$$D_{G,dG}(r, r') = \mathfrak{P}(\varphi(r), p_r) - \mathfrak{P}(\varphi(r'), p_r), \tag{4.11}$$

$$\mathfrak{P}(q, o) = G(\phi(o)) - D_{G,dG}(\phi(o), \varphi^{-1}(q)) \tag{4.12}$$

for all $q \in \mathcal{Q}'$, where $G = C^*$ and dG is a subgradient of G, and $\mathbb{E}_{p_r}[\phi] = r$ for all $r \in \mathcal{R}'$.

Note that the expression (4.11) is the net profit of a single trade in the standard prediction market framework. That is, the right-hand side may be written $\langle \phi(o), q' - q \rangle - (C(q') - C(q))$, which is just the payoff of the trade $q' - q$ minus its cost. Hence, Corollary 4.6 says something surprising: the net profit of a trade in a prediction market can be expressed as the divergence between the prices before and after the trade. This can be useful when designing a prediction market according to some notion of "information distance" — one can give the traders a direct incentive to minimize the distance between the market price and their belief, for any *Bregman* information distance.

Finally, to illustrate why extra care is needed for mutual reducibility, we give a very simple example of a nondifferentable G for which condition (4.8) and Theorem 4.3 do not hold. In this case, the prediction market has strictly more expressiveness (in terms of the possible payoffs) than the corresponding scoring rule. The G here is the same as in Example 3.1 and Figure 3.1.

Example 4.1. *Let $G(r) = |r| + r^2/2$ and $dG_r = r + \operatorname{sgn}(r)$, where $\operatorname{sgn}(r)$ denotes the sign of r (which is 0 at $r = 0$). By a simple computation, $C(q) = G^*(q) = (q - \operatorname{sgn}(q))^2/2$ when $|q| > 1$ and $C(q) = 0$ on $[-1, 1]$. Now let $q = 1/2$, which has $\partial C_q = \{0\}$. By condition (4.8) we would need $0 = \langle q - dG_0, 0 - \phi(o) \rangle = q \cdot \phi(o)$ for all o. Thus, as long as $\phi(o) \neq 0$ for some o, we have violated eq. (4.8).*

Moreover, we clearly cannot reduce $\mathfrak{P}$ to $\mathfrak{S}$ in this case, since each $q \in [-1, 1]$ would have to have $\varphi(q) = 0$, but except for $q = 0$ and $\phi(o) = 0$, we have $\langle q, \phi(o) \rangle \neq 0 = \langle dG_{\varphi(q)}, \phi(o) \rangle$. In essence, the prediction market is much more expressive with regard to the belief $\mathbb{E}_p[\phi] = 0$, offering agents an infinite array of utility functions, whereas the scoring rule offers only one.

4.3 Generalized exponential families

As mentioned in §4.1.3, *generalized exponential familias (GEFs)* are an extension of classical[4] exponential families as maximum entropy distributions for non-standard entropy functions. Below we take the reader through the very basics of the exponential family derivation, and show how one may generalize them in a natural[5] way for other choices of entropy.

The goal of this section is two-fold. First, we wish to develop the theory of generalized exponential families beyond the excellent foundation of Grünwald and Dawid [54]. Second, and more relevant to the broader story of this chapter, we seek the machinery necessary to relate GEFs to divergences, scoring rules, and prediction markets. In many cases, the relationships we uncover are surprising; as it happens, GEFs are the "answer" to many natural questions one may ask about the other constructs.

As before, we assume that $\mathcal{P}$ is a convex set of distributions. Throughout this section and the next we will write $\phi^\top \theta$ to mean $o \mapsto \langle \phi(o), \theta \rangle$. We will also be assuming $\mathcal{P} \subseteq \mathcal{W}$ for the dual pair $(\mathcal{W}, \mathcal{W}^*)$ given by the duality $\langle p, f \rangle = \mathbb{E}_p[f]$, as mentioned in the introduction.

[4] Throughout, we use the term "classical" to mean the standard definition in the literature.

[5] Though in same cases it may be more natural to be mean than natural; see p. 77.

Note: there are many symbols and spaces at play in this section, and for those unfamiliar with exponential families (and even more for those who are, as we use somewhat nonstandard notation), it may be helpful to refer to Figure 4.3 as a roadmap.

Our techniques are almost exclusively grounded in convex analysis. In particular, we will assume for the whole section that we are working with proper and lower semi-continuous (abbreviated l.s.c.; see Definition 3.7) convex functions f, so that the Fenchel–Moreau Theorem applies and we may conclude that $f^{**} = f$. Beyond this, a major source of seemingly mysterious results is the simple observation that by linearity we may pass between inner products in (W, W^*) to those in $(\mathcal{V}, \mathcal{V}^*)$. Specifically, we will constantly appeal to the fact that $\langle p, \phi^\top \theta \rangle = \langle \mathbb{E}_p[\phi], \theta \rangle$.

Exponential families

Before formally introducing generalized exponential families, we recall some basic concepts from the literature on classical exponential families. In this case, we have a some base measure ν on $(\mathcal{O}, \Sigma)$, and a statistic $\phi : \mathcal{O} \to \mathbb{R}^k$, and a parameter space Θ called the *natural parameters*. The exponential family $\{p_\theta\}_\Theta$ is defined by

$$p_\theta(o) = \exp\{\phi(o)^\top \theta - \Psi(\theta)\}, \tag{4.13}$$

where $\Psi(\theta)$ is chosen to normalize p_θ, namely

$$\Psi(\theta) = \log \int_{\mathcal{O}} \exp\{\phi(o)^\top \theta\} \, d\nu(o). \tag{4.14}$$

Typically the parameter space Θ is actually defined in terms of Ψ, letting $\Theta \doteq \{\theta \in \mathbb{R}^k : \Psi(\theta) < \infty\}$.

Many interesting characteristics of exponential families are known (see e.g. [18, 98]), but for our exploration two are especially relevant: exponential families have alternate parameterizations in terms of the mean of the statistic ϕ, and they can also be viewed as maximum entropy distributions under a mean constraint. Very briefly, one can check that, surprisingly, $\nabla \Psi(\theta) = \mathbb{E}_{p_\theta}[\phi]$; that is, the derivative of Ψ at θ is precisely the ϕ-mean for p_θ. This allows one to reparametrize the family by $\mathbb{E}_{p_\theta}[\phi]$. Moreover, one can derive this mean parametrization via a maximum entropy calculation. We briefly sketch this argument now.

The widely-used notion of entropy in probability theory is that of *Shannon entropy*, defined as

$$H(p) = -\int_{\mathcal{O}} p(o) \log p(o) \, d\nu(o). \tag{4.15}$$

The principle of maximum entropy states that given some data with empirical mean $\hat{\mu}$, to estimate the distribution from which the data was generated, one should compute the distribution p of maximum entropy $H(p)$ under the constraint $\mathbb{E}_p[\phi] = \hat{\mu}$. Formally, we wish to perform the following optimization.

$$p \in \operatorname{argsup}\{H(p) : p \in \mathcal{P}, \mathbb{E}_p[\phi] = \hat{\mu}\}. \tag{4.16}$$

To solve this problem, we may turn to variational analysis and Lagrange multipliers, yielding the solution (4.13) [38, Thm 12.1.1], where θ is the vector of Lagrange multipliers from the calculation.

One can also derive (4.13) via convex analysis. As it turns out, Ψ is a convex function, and happens to equal $(-H)^*$, the convex conjugate of (negative) Shannon entropy, applied to a particular point. To see this, first compute the entropy dual,

$$(-H)^*(q) = \log \int_{\mathcal{O}} \exp\{q(o)\}\, d\nu(o), \tag{4.17}$$

and from there we can check that indeed $\Phi(\theta) = (-H)^*(\phi^\top \theta)$. More surprisingly, we can rederive (4.13) via the *derivative* of $(-H)^*$ at the same point:

$$\nabla_q (-H)^*(o) = \frac{\exp\{q(o)\}}{\log \int_{\mathcal{O}} \exp\{q(o')\}\, d\nu(o')} = \exp\{q(o) - H^*(q)\}, \tag{4.18}$$

whence we have

$$\nabla_{\phi^\top \theta} (-H)^*(o) = \exp\{\phi(o)^\top \theta - \Psi(\theta) = p_\theta(o). \tag{4.19}$$

As we will see below, this derivation via convex analysis is in essence the same as the maximum entropy calculation.

As a final note, it is common in the literature to restrict to *regular* families, defined as follows.

Definition 4.7. *An exponential family $\{p_\theta\}_\Theta$ as defined by eq. (4.13) is* regular *if Θ is an open set.*

Generalizing to other entropies

As we will show in this section, all of the observations and properties of classical exponential families mention above can be extended to families derived from other entropy functions other than Shannon entropy. We are heavily influenced by Grünwald and Dawid [54], who introduced idea of generalized exponential families, along with many of the ideas we will explore. Our approach is different, however, relying much more on convex analysis. As a result, our setting will be slightly less general, but the extra regularity will go a long way. See the discussion before §4.3.1 for more details.

We first set limits on what we will consider an alternate entropy. NB: our entropy functions will always be *convex*, unlike Shannon entropy which is concave; the reader may need to mentally insert negations if more comfortable with the latter.

Definition 4.8. *A function $F : \mathcal{P} \to \mathbb{R}$ is a* (generalized) entropy function *if it is convex, l.s.c., and proper.*

Recall that we adopt the convention $F(x) = \infty$ for $x \notin \mathsf{dom}(F)$, i.e., $F(p) = \infty$ for all $p \notin \mathcal{P}$. We can now define our generalized version of an exponential family. Of all

the derivations above for classical families, the final derivation (4.19) lends itself most to generalization. We employ this strategy, replacing $-H$ with our alternate entropy F.

Definition 4.9. *Let F be a given generalized entropy function, and let statistic $\phi : \mathcal{O} \to \mathbb{R}^k$ be given. Then a family of distributions $P_\Theta = \{p_\theta \in \partial F^*(\phi^\top \theta)\}_{\theta \in \Theta}$ is a* F-generalized exponential family (F-GEF) *with* cumulant $C(\theta) \doteq F^*(\phi^\top \theta)$, *where* $\Theta \doteq \mathsf{dom}(C)$.

Of course, we have not yet shown that indeed $P_\Theta \subseteq \mathcal{P}$. To see this, apply Corollary 3.14 to any $d \in \mathsf{dom}(F^*)$; then for all $w \in \mathcal{W}$, we have $w \in \partial F_d^* \implies d \in \partial F_w$. Since $\partial F_w = \emptyset$ for $w \notin \mathsf{dom}(F)$, we must have $\partial F^* \subseteq \mathsf{dom}(F) = \mathcal{P}$. In particular then, $p_\theta \in \mathcal{P}$ for all $\theta \in \Theta$.

Definition 4.10. *The F-GEF $\{p_\theta\}_{\theta \in \Theta}$ is* regular *if its cumulant C is l.s.c. and proper.*

It is a classic result in convex analysis that given convex f and a linear map A, the function $Af \doteq x \mapsto \inf\{f(y) : Ax = y\}$ is convex if f is, and satisfies $(Af)^* = f^* \circ A^\top$. We state and prove this result specifically for our setting; see e.g. [95, 103] for more depth.

Lemma 4.7. *Let $G(v) \doteq \inf\{F(p) : p \in \mathcal{P}, \mathbb{E}_p[\phi] = v\}$. Then G is convex, and $G^*(\theta) = F^*(\phi^\top \theta)$.*

Proof. Note that $\mathcal{R} \doteq \mathsf{dom}(G) = \{\mathbb{E}_p[\phi] \,|\, p \in \mathcal{P}\}$, by the convention $\inf \emptyset = \infty$. This $\mathcal{R}$ is convex by convexity of $\mathcal{P}$ and linearity of $\mathbb{E}[\phi]$. We first show convexity of G.

$$\begin{aligned}
G(\alpha r + (1-\alpha)r') &= \inf_{p \in \mathcal{P}} \{F(p) : \mathbb{E}_p[\phi] = \alpha r + (1-\alpha) r'\} \\
&\leq \inf_{p, p' \in \mathcal{P}} \{F(\alpha p + (1-\alpha)p') : \mathbb{E}_p[\phi] = r,\, \mathbb{E}_{p'}[\phi] = r'\} \\
&\leq \inf_{p, p' \in \mathcal{P}} \{\alpha F(p) + (1-\alpha) F(p') : \mathbb{E}_p[\phi] = r,\, \mathbb{E}_{p'}[\phi] = r'\} \\
&= \alpha G(r) + (1-\alpha) G(r').
\end{aligned}$$

We now compute G^* directly.

$$\begin{aligned}
G^*(\theta) &= \sup_{r \in \mathcal{R}} \langle r, \theta \rangle - G(r) \\
&= \sup_{r \in \mathcal{R}} \langle r, \theta \rangle - \inf_{p \,:\, \mathbb{E}_p[\phi] = r} F(p) \\
&= \sup_{r \in \mathcal{R}} \sup_{p \,:\, \mathbb{E}_p[\phi] = r} \langle r, \theta \rangle - F(p) \\
&= \sup_{p \in \mathcal{P}} \left\langle p, \phi^\top \theta \right\rangle - F(p) \\
&= F^*(\phi^\top \theta),
\end{aligned}$$

where the penultimate equality follows by $\left\langle p, \phi^\top \theta \right\rangle = \langle \mathbb{E}_p[\phi], \theta \rangle$ and by definition of $\mathcal{R}$. □

Thus, Lemma 4.7 implies that the cumulant C of an F-GEF is the convex conjugate of $G(v) \doteq \inf\{F(p) : p \in \mathcal{P}, \mathbb{E}_p[\phi] = v\}$. It is interesting to compare this result to our review of classical exponential families above; there we had $C = \Psi$, and as we saw, $\Psi(\theta) = (-H)^*(\phi^\top \theta)$, a result we of course recover by setting $F = -H$. We now show a generalization of another classical result, that $\nabla\Psi(\theta) = \mathbb{E}_{p_\theta}[\phi]$. Note that because of our generality, our entropy might not be differentiable in any sense, so we instead fall back on subgradients.

Proposition 4.8. *A regular F-GEF $\{p_\theta\}$ with statistic ϕ and cumulant C satisfies $\mathbb{E}_{p_\theta}[\phi] \in \partial C_\theta$ for all θ.*

Proof. First we prove a small claim:

> **Claim.** Let $f : X \to \mathbb{R}$ convex and $A : X \to Y$ linear, and set $g \doteq f \circ A^\top$. Then $d \in \partial f(A^\top y) \implies Ad \in \partial g(y)$.
> *Proof.*
> $$\begin{aligned} &\forall x\ f(x) \geq f(A^\top y) + \left\langle d, x - A^\top y \right\rangle \\ \implies &\forall y'\ f(A^\top y') \geq f(A^\top y) + \left\langle d, A^\top y' - A^\top y \right\rangle \\ \iff &\forall y'\ g(y') \geq g(y) + \left\langle Ad, y' - y \right\rangle . \end{aligned}$$

Applying this to $f = F^*$, $g = C$, $A : p \mapsto \mathbb{E}_p[\phi]$ (and thus $A^\top = \phi^\top$) yields $p \in \partial F^*(\phi^\top \theta) \implies \mathbb{E}_p[\phi] \in \partial C(\theta)$, from which the result follows. □

We now see why we are justified in calling C the *cumulant*, as it generates the first moment of the statistic. In fact, in light of Proposition 4.8, we are justified to follow the classical case and identify each distribution p_θ with its ϕ-mean. In this way we may parametrize P_Θ (or a subset thereof; see below for discussion) by the means.

Definition 4.11. *Given F-GEF $P_\Theta = \{p_\theta\}_{\theta \in \Theta}$, we define the* mean parameters *to be $\mathcal{R}_\Theta = \{\mathbb{E}_{p_\theta}[\phi]\}_{\theta \in \Theta}$, and a* mean parameterization *of P_Θ to be a family $\{p_r\}_{r \in \mathcal{R}_\Theta} \subseteq P_\Theta$ such that $\forall r \in \mathcal{R}_\Theta$, $\mathbb{E}_{p_r}[\phi] = r\}$.*

Note that in general, there may be multiple distributions in P_Θ with the same mean, and $\mathcal{R}_\Theta$ may be non-convex. However, by Proposition 4.8, we see that there is a unique mean parametrization if C is strictly convex, since $\mathbb{E}_{p_\theta}[\phi] \in \partial C(\theta)$ but strict convexity implies that the sets $\partial C(\theta)$ are disjoint. Moreover, if C is continuously differentiable, then $\mathbb{E}_{p_\theta}[\phi] \in \partial C(\theta) = \{\nabla C(\theta)\}$, so $\mathcal{R}_\Theta$ is convex as the range of $\mathcal{P}$ under linear map $p \mapsto \mathbb{E}_p[\phi]$; in other words, every mean of ϕ is realized in P_Θ. While uniqueness of the parametrization and convexity of $\mathcal{R}_\Theta$ are certainly desirable properties, our results will not rely on these assumptions.

We now return to our above discussion and sketch how to derive a GEF $\{p_\theta\}_\Theta$ as the maximum entropy family under a mean constraint. We may define $p_r \in \operatorname{arginf}\{F(p) : p \in \mathcal{P}, \mathbb{E}_p[\phi] = r\}$, which when the arginf is nonempty for all r, is a parameterized family of distributions. But is $P_\mathcal{R} \doteq \{p_r\}_\mathcal{R}$ a mean parameterization of an F-GEF? As we will see

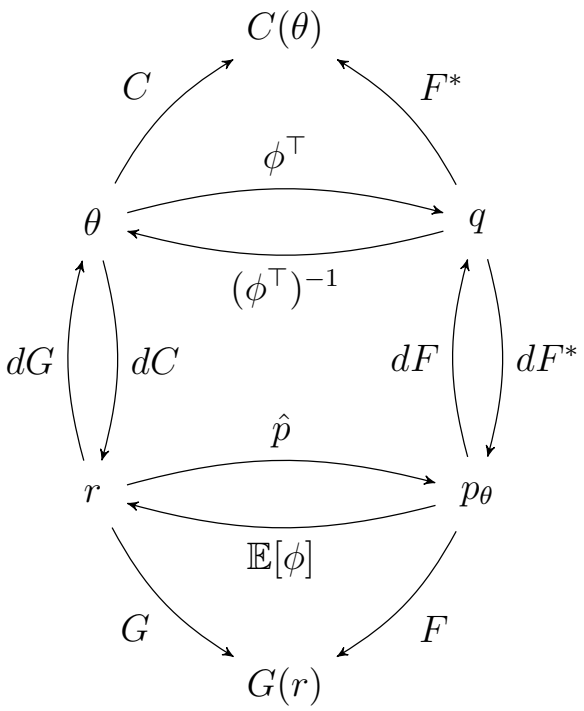


Figure 4.2: An illustration of the relationship between the various spaces, symbols, and maps in this section. Here $(\phi^\top)^{-1}$ denotes the left inverse of $\phi^\top$. Note that the diagram is not necessarily commutative, though starting from r and θ it is commutative for the most part when when F and F^* are strictly convex and differentiable.

in the proof of Theorem 4.14, and the discussion thereafter, we may select $P_{\mathcal{R}}$ such that $\phi^\top dG_r \in \partial F(p_r)$, where G is defined as in Lemma 4.7. But then Corollary 3.14 gives $p_r \in \partial F^*(\phi^\top dG_r)$, which means indeed $p_r = p_{\theta(r)}$ for an F-GEF P_Θ.

Finally, we remark about the relationship between our approach and that of Grünwald and Dawid [54]. In [54, § 7.4], GEFs are *defined* by the maximum entropy calculation discussed above, and hence as long as the arginf is nonempty, their results go through. After defining what we call the mean parametrization, they introduce the "natural" parameterization in eq. (53), essentially as the argsup in the calculation of F^*. While in many senses, the results of [54] hold more generally than ours, we choose instead to restrict to the l.s.c. and proper setting, so that we may start with the natural parameters, and use them as our fundamental definition. This allows us to draw closer connections to the existing theory of exponential families, but as we will see, it also allows us to weave our results into our greater story involving divergences, scoring rules, and prediction markets.

4.3.1 Divergences

We now apply the above definitions and machinery to show a bijection between generalized exponential families and generalized Bregman divergences. This investigation is inspired by Banerjee et al. [16], which shows a similar result for (classical) exponential families. In fact, our result will in some sense generalize theirs to other entropies besides Shannon entropy.

Definition 4.12. *A generalized Bregman divergence $D_{G,dG}$ for a convex $G : \mathcal{V} \to \overline{\mathbb{R}}$ is F-*regular *for a convex $F : \mathcal{P} \to \overline{\mathbb{R}}$ if G is proper and l.s.c., and there exists some statistic $\phi : \mathcal{O} \to \mathbb{R}$ such that*

$$G(v) = \inf_{p \in \mathcal{P}} \{F(p) : \mathbb{E}_p[\phi] = v\}. \tag{4.20}$$

We will also say G itself is F-regular with statistic ϕ.

The bijection we will show will tie Bregman divergences to certain *equivalence classes* of GEFs, which we now define.

Definition 4.13. *The* cumulant class *of F-GEFs with cumulant C is the set of F-GEFs whose cumulant is C.*

Of course, two F-GEFs with the same cumulant may be very different from one another, but in some sense the disparity is only with regard to $\mathcal{O}$, not Θ and $\mathcal{R}$; see the "Overlapping bijections" discussion below. Note that our definition of regularity, Definition 4.10, really is a property of the cumulant, and hence naturally applies to cumulant classes; we will say a cumulant class is regular to mean its cumulant satisfies the same.

Theorem 4.9. *Fix entropy function F. The set of F-regular Bregman divergences is in bijection with the set of regular cumulant classes of F-GEFs.*

Proof. Specifically, we will show that each F-regular Bregman divergence $D_{G,dG}$ yields the cumulant class of $C(\theta) \doteq G^*(\theta)$, and as the convex conjugate is invertible, we will thus establish our bijection.

Let F-regular Bregman divergence $D_{G,dG}$ be given, with statistic ϕ. Then simply take the cumulant class of the F-GEF $P_\Theta = \{p_\theta \in \partial F^*(\phi^\top \theta)\}_{\theta \in \Theta}$, where $\Theta = \mathsf{dom}(G^*)$. By Lemma 4.7, we have that the cumulant of P_Θ is G^*. Finally, by regularity of $D_{G,dG}$ and Theorem 3.13, we have $(G^*)^{**} = (G^{**})^* = G^*$, so P_Θ is regular.

Now, given any cumulant class of regular F-GEFs with cumulant C and statistic ϕ, we take the Bregman divergence with convex function $G \doteq C^*$. Again by regularity and Theorem 3.13, we have that G is l.s.c. and proper. Also, since $G^* = C^{**} = C$, we have $G^*(\theta) = F^*(\phi^\top \theta)$, so by Lemma 4.7 we must have $G(v) = \inf_{p\,:\,\mathbb{E}_p[\phi]=v} F(p)$, meaning G is F-regular. □

We now try to consider the result of Banerjee et al. [16]. They give a bijection between a certain class of classical exponential families and a certain class of Bregman divergences, so to begin we state generalized versions of their conditions. We will use the concept of

Legendre type from convex analysis (cf. [89, §26]) which is the following property of a set and function pair (X, F): (a) X is nonempty and open, (b) F is strictly convex and differentiable on X, and (c) $\lim_{x\to b} \|\nabla f(x)\| = \infty$ for $x \in X$ and $b \in \mathrm{bd}(X)$, the boundary of X.

Definition 4.14. *A F-GEF P_Θ with cumulant C is* super-regular *if it is regular and additionally $\Theta = \mathrm{int}(\mathsf{dom}(C))$ and (Θ, C) is of Legendre type.*

Definition 4.15. *A Bregman divergence D_G is F-*super-regular *if it is regular and additionally $(\mathrm{int}(\mathsf{dom}(G)), G)$ is of Legendre type.*

We briefly sketch why these regularity conditions roughly correspond to those of Banerjee et al. when $F = -H$. Their condition on divergences D_G is that $G = C^*$ for some strictly convex C which (after applying a result due to Devinatz) is of the form $C(\theta) = \log \int_{\mathbb{R}^k} \exp\{\langle x, \theta\rangle\} d\mu(x)$ for some unique bounded non-negative measure μ. Of course, applying Lemma 4.7 and appealing to eq. (4.18), we easily see that $-H$-super-regular implies their condition. Exponential family regularity follows much more simply, as an application of Lemma 1 of Banerjee et al. [16].

We now state a refinement of Theorem 4.9, for the super-regular case. When $F = -H$, and the statistic ϕ is minimal (affinely independent), we obtain essentially the same bijection as [16]. The proof follows from a classic result of convex analysis, that (X, F) is of Legendre type if and only if (X^*, F^*) is [89, Thm 26.5].

Corollary 4.10. *Fix entropy function F. The set of F-super-regular Bregman divergences is in bijection with the set of cumulant classes of F-GEFs.*

It has been noted that one can express the relative entropy between two members of a classical exponential family as a divergence between their corresponding parameters [6, 23, 76]. Specifically, one can write

$$\mathrm{KL}(p_\theta \| p_{\theta'}) = D_\Psi(\theta', \theta). \tag{4.21}$$

We would like to generalize this result to our setting, hopefully to show that the *generalized relative entropy*, given by D_F, can be written analogously. First we prove a useful lemma.

Lemma 4.11. *Let regular F-GEF $\{p_\theta\}$ and F-regular G be given, both with statistic ϕ. Then $F(p_\theta) = G(\mathbb{E}_{p_\theta}[\phi])$ for all θ.*

Proof. By Proposition 4.8, we have $\mathbb{E}_{p_\theta}[\phi] \in \partial C(\theta)$, which by Lemma 3.11 implies that $C^*(\mathbb{E}_{p_\theta}[\phi]) = \langle \mathbb{E}_{p_\theta}[\phi], \theta\rangle - C(\theta)$. By Lemma 4.7, and the fact that $G^{**} = G$, we have $C^* = G$. Moreover, $\langle \mathbb{E}_{p_\theta}[\phi], \theta\rangle = \left\langle p_\theta, \phi^\top \theta \right\rangle$, and $C(\theta) = F^*(\phi^\top \theta)$, so we now have $G(\mathbb{E}_{p_\theta}[\phi]) = \left\langle p_\theta, \phi^\top \theta \right\rangle - F^*(\phi^\top \theta)$. Finally, as $p_\theta \in \partial F^*(\phi^\top \theta)$ by definition, applying Lemma 3.11 once more yields $G(\mathbb{E}_{p_\theta}[\phi]) = \left\langle p_\theta, \phi^\top \theta \right\rangle - F^*(\phi^\top \theta) = F(p_\theta)$. □

We now generalize equation (4.21) to F-GEFs.

Proposition 4.12. *Let F-GEF $\{p_\theta\}$ be given with cumulant C. Then there exist subradients dF and dC such that for all $\theta, \theta' \in \Theta$,*

$$D_{F,dF}(p_{\theta'}, p_\theta) = D_{C,dC}(\theta, \theta').$$

Proof. By Lemma 3.11, and the fact that $p_\theta \in \partial F^*(\phi^\top \theta)$ by definition, we have $F^*(\phi^\top \theta) = \langle p_\theta, dF_{p_\theta} \rangle - F(p_\theta)$ for all $\theta \in \Theta$ and any choice of subgradient dF. Thus, selecting $dF_{p_\theta} = \phi^\top \theta$, we have

$$\begin{aligned} D_{F,dF}(p_{\theta'}, p_\theta) &= F(p'_\theta) - F(p_\theta) - \left\langle p_{\theta'} - p_\theta, \phi^\top \theta \right\rangle \\ &= \left\langle p_{\theta'}, \phi^\top \theta' \right\rangle - F^*(\phi^\top \theta') - \left\langle p_\theta, \phi^\top \theta \right\rangle + F^*(\phi^\top \theta) \\ &\quad - \left\langle p_{\theta'} - p_\theta, \phi^\top \theta \right\rangle \\ &= F^*(\phi^\top \theta) - F^*(\phi^\top \theta') + \left\langle p_{\theta'}, \phi^\top \theta' - \phi^\top \theta \right\rangle \\ &= C(\theta) - C(\theta') + \left\langle \mathbb{E}_{p_{\theta'}}[\phi], \theta' - \theta \right\rangle, \end{aligned}$$

where we used Lemma 4.7 in the last step. Finally, by Proposition 4.8 we may select $dC_\theta = \mathbb{E}_{p_\theta}[\phi]$, completing the proof. □

We may also derive a similar result for a mean parameterization.

Proposition 4.13. *Let entropy F be given, and let G be F-regular with statistic ϕ, with G additionally being strictly convex with subgradient dG. Then there exists a subgradient dF such that for any mean-parametrized F-GEF $P_\mathcal{R} = \{p_r\}_{r \in \mathcal{R}}$ with statistic ϕ, and any $r, r' \in \mathcal{R}$,*

$$D_{F,dF}(p_r, p_{r'}) = D_{G,dG}(r, r').$$

Proof. We first show that for any set of distributions $\{p_r\}_r$ with (1) $\mathbb{E}_{p_r}[\phi] = r$ and (2) $G(r) = F(p_r)$, the result holds with $dF_{p_r} \doteq \phi^\top dG_r$:

$$\begin{aligned} D_{F,dF}(p_r, p_{r'}) &= F(p_r) - F(p_{r'}) - \left\langle p_r - p_{r'}, \phi^\top dG_r \right\rangle \\ &= G(r) - G(r') - \left\langle \mathbb{E}_{p_r}[\phi] - \mathbb{E}_{p_{r'}}[\phi], dG_r \right\rangle \\ &= D_{G,dG}(r, r'). \end{aligned}$$

Our set $P_\mathcal{R}$ trivially satisfies (1); for (2), note that by Definition 4.11 we have some $\theta(r)$ such that $p_r = p_{\theta(r)}$. Then by Lemma 4.11, $G(r) = G(\mathbb{E}_{p_{\theta(r)}}[\phi]) = F(p_{\theta(r)}) = F(p_r)$. □

Propositions 4.13 and 4.12 give life to our bijection in Theorem 4.9. From them we see that not only are F-GEFs in bijection with divergences, but these divergences exactly capture the geometry of the GEF, and succinctly so, in terms of their parameters.

Other remarks

The above exploration of generalized exponential families certainly opens more doors than it closes. It is in particular quite natural to ask, for each succinct quality of classical exponential families, does a generalization of this quality hold for GEFs? We briefly touch on a few of these questions and other points.

Sufficiency. It is well known that the statistic ϕ is sufficient for a classical exponential family, meaning that the likelihood of some data $\{o_i\}_i$ depends on the underlying parameter θ only through the empirical mean of the statistic, specifically $\hat{\mu} = \frac{1}{n}\sum_i \phi(o_i)$. This property follows directly from eq. (4.13). For an F-GEF, though, this property does not hold in general, simply because $\nabla F^*(\phi^\top\theta)_o \neq h(o)g(\theta, \phi(o))$. Another notion of sufficiency does hold, however: for any loss function $L : \mathcal{R}\times\mathcal{O} \to \mathbb{R}$ which is proper for the map $\Gamma(p) = \mathbb{E}_p[\phi]$, the loss of prediction r depends on the data only through $\hat{\mu}$. This follows from Theorem 4.1, simply negating to turn the score into a loss.

Conjugate priors. One of the most useful properties of classical exponential families is the existence of *conjugate priors*, families of prior distributions which when conditioned on data from an exponential family yield a posterior from the same family as the prior. It is not clear whether analogous families of priors exist for generalized exponential families. An interesting direction to explore would be to find a new notion of Bayesian updating for GEFs which captured the "geometry" of the entropy function F, and then use this updating to define conjugate priors.

Overlapping bijections. The observant reader will note that there is a complicated relationship between entropies F and the possible F-regular convex functions G. In particular, while it seems clear that the same F yields many different F-regular functions G by changing ϕ, it is not clear whether there exists an F which can yield all possible G's. The converse relationship is also interesting: a G can be F-regular for many F's. Hence, there is a delicate many-many relationship between F and G. We briefly illustrate with some examples.

Working with $\mathcal{P} = \Delta_n$ for simplicity, take $F(p) = \|p\|^2/2$. It is easy to see that any G which is F-regular must be piecewise-quadratic. (In the regime where p_r has all positive entries, this follows from $C(\theta) = F^*(\phi^\top\theta) = \|\phi^\top\theta\|^2/2$, so $G = C^*$ must also be quadratic; similar reasoning applies to the nonzero entries once p_r hits the edge of Δ_n.) Hence, even when $k = \dim(\Theta) \ll n$, G will always be piecewise quadratic when $F = \|\cdot\|^2$.

Conversely, it is trivial to come up with different functions F yielding the same G. The easiest way to do this is to just relabel the ground set $\mathcal{O}$ (and ϕ accordingly), so that F has changed due to a permutation of the arguments, but the optimization yielding G remains equivalent. A less superficial method, given any G and ϕ, is to take the function $F(p) = G(\mathbb{E}_p[\phi]) + f(p - \hat{p}(\mathbb{E}_p[\phi]))$, where $\hat{p}$ is a continuous inverse of ϕ and f is positive definite.[6] Of course, we must restrict to choices of $\hat{p}$ and f which keep F convex.

Finally, we point out that, like the quadratic example, taking $F = -H$ is very restrictive in terms of the possible functions G that are F-regular. In particular, it is easy to see that no $-H$-regular G can be piecewise quadratic: $C(\theta) = \log\sum_i e^{\theta_i}$ has nonzero derivatives of all orders, but $C = G^*$ would have locally quadratic points, with zero derivatives beyond order 2. Thus, the notion of "regular" used by Banerjee et al. [16], which essentially captures "log-exp-convex" functions, excludes a very natural class of convex functions.

[6] A function f is positive definite if $f(0) = 0$ and $f(x) > 0$ for all $x \neq 0$.

4.3.2 Inducing scoring rules

Someone charged with the task of designing a scoring rule $\mathfrak{S}$ to elicit the mean of some statistic might naturally wonder whether it suffices to just select a classical scoring rule $\mathfrak{S}^{\mathcal{P}}$ defined on $\mathcal{P}$ and use that — that is, can a designer simply *induce* $\mathfrak{S}$ from $\mathfrak{S}^{\mathcal{P}}$? Specifically, letting $\Gamma : p \mapsto \mathbb{E}_p[\phi]$ denote our linear property as usual, it seems natural to pick an arbitrary $\hat{p} : \mathcal{R} \to \mathcal{P}$ mapping means r to some $\hat{p}(r) \in \Gamma_r$,[7] and take

$$\mathfrak{S}(r, o) \doteq \mathfrak{S}^{\mathcal{P}}(\hat{p}(r), o). \tag{4.22}$$

As we will see, this can be done, but care must be taken in selecting $\hat{p}$. Even assuming that equation (4.22) can be satisfied, other questions remain:

1. Does one lose design flexibility by restricting to scores $\mathfrak{S}$ of the form (4.22)?
2. Are there multiple choices of $\mathfrak{S}^{\mathcal{P}}$ and $\hat{p}$ that yield the same $\mathfrak{S}$?
3. Fixing $\mathfrak{S}^{\mathcal{P}}$, what scoring rules $\mathfrak{S}$ can be obtained by varying $\hat{p}$ or ϕ?

We will answer many of these questions in this section, as well as similar questions for prediction markets.

To begin, let us see why not all choices of $\hat{p}$ suffice for (4.22). Working with $\mathcal{P} = \Delta_n$, take $\mathfrak{S}^{\mathcal{P}}$ to be the Brier score $\mathfrak{S}^{\mathcal{P}}(p, o) = 2p_o - \|p\|^2$. Now for statistic $\phi : o \mapsto \mathbb{1}\{o = 1\}$, the indicator for $o = 1$, we will try $\hat{p}(r) = [r, 1 - r, 0, \ldots, 0] \in \mathcal{P}$, the distribution with mass r on 1, $1 - r$ on 2, and 0 otherwise. Clearly $\mathbb{E}_{\hat{p}(r)}[\phi] = r$, but given $p = [p_1, p_2, \ldots, p_n]$, solving

$$r(p) = \operatorname*{argsup}_r \mathfrak{S}^{\mathcal{P}}(\hat{p}(r), p) = \operatorname*{argsup}_r \|p - \hat{p}(r)\|^2, \tag{4.23}$$

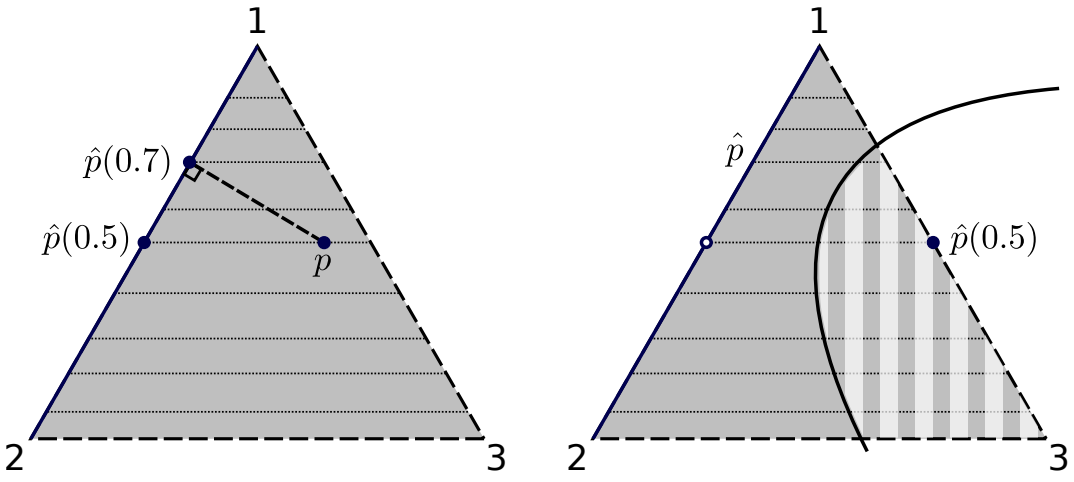
which we can compute by solving $0 = \nabla_r(2p_1 r + 2p_2(1 - r) - r^2 - (1 - r)^2)$, yields $r(p) = (p_1 - p_2 + 1)/2$. But now we see $r(p) \neq \mathbb{E}_p[\phi]$, and hence $\mathfrak{S}(r, o) = \mathfrak{S}^{\mathcal{P}}(\hat{p}(r), o)$ is not proper! See Figure 4.3 for an illustration.

For a more extreme example, one can take $\hat{p}(r)$ the same as above but with $\hat{p}(1/2) = [1/2, 0, 1/2, 0, \ldots, 0]$. As we saw in eq. (4.23), the geometry exhibited by the Brier score is Euclidean, so clearly any p in a small enough ball around $\hat{p}(1/2)$ would prefer $r = 1/2$ to the correct report. In fact, as Figure 4.3 shows, the region of points p for which the agent will report $r = 1/2$ is the interior of an entire parabolic region of the simplex.

In general, it seems that our choice $\hat{p}$ must somehow conform to the geometry of $F(p) \doteq \mathfrak{S}^{\mathcal{P}}(p, p)$. In particular, it must have the property that for all $p : \mathbb{E}_p[\phi] = r$, the "closest" point in $\{\hat{p}(r') : r' \in \mathcal{R}\}$ to p is $\hat{p}(r)$, where distance is measured by $\mathfrak{S}^{\mathcal{P}}$. In fact, there are interesting connections to information geometry (cf. Amari [6]), where one may view $\hat{p}$ as a kind of geodesic, though we do not explore this further here.

We now address the central question: given a linear property $\Gamma : p \mapsto \mathbb{E}_p[\phi]$, can one induce a Γ-proper scoring rule $\mathfrak{S}$ from a classical $\mathfrak{S}^{\mathcal{P}} : \mathcal{P} \times \mathcal{O} \to \overline{\mathbb{R}}$? The following result

[7] Recall that $\Gamma_r \doteq \{p \in \mathcal{P} : \Gamma(p) = r\}$.



(a) (b)

Figure 4.3: Two 3-outcome examples of ways to choose $\hat{p}$ which do not yield a proper scoring rule for Γ when $\mathfrak{S}^{\mathcal{P}}(p, p) = \|p\|^2$. The horizontal dotted lines depict a selection of level sets of Γ. In (a), an agent with belief p would report 0.7 instead of $\Gamma(p) = 0.5$. In (b), the striped region is the set of points p for which the agent would report $r = 0.5$.

answers this in the affirmative, and also addresses some of our other questions. Before stating the result, we introduce two simple concepts. The first is the *Bayes risk* from decision theory, which roughly speaking is the score corresponding to the worst-case underlying distribution consistent with the report r. The second is a technical condition which ensures that our mean parameterizations have no "holes." We will also again use the notion of a *fair* scoring rule; see Definition 4.4.

Definition 4.16. *The* Bayes risk *of scoring rule $\mathfrak{S} : \mathcal{R} \times \mathcal{O} \to \overline{\mathbb{R}}$ with respect to property Γ is defined by* $\mathsf{Risk}(\mathfrak{S}) \doteq (r \mapsto \inf_{p \in \Gamma_r} \mathfrak{S}(r, p))$.

Definition 4.17. *A convex function F is Γ-bounded for a property Γ if F achieves its lower bound on Γ_r for all $r \in \mathsf{relint}(\mathcal{R})$. That is, if for all $r \in \mathsf{relint}(\mathcal{R})$,* $\operatorname{arginf}_{p \in \Gamma_r} F(p) \neq \emptyset$.

In the following theorem, we say "G is F-regular on S" to mean there exists $\Gamma : p \mapsto \mathbb{E}_p[\phi]$ such that $G(r) = \inf_{p \in \Gamma_r} F(p)$ for all $r \in S$.

Theorem 4.14. *Let $\Gamma : p \mapsto \mathbb{E}_p[\phi]$ be a linear map, and let Γ-fair, Γ-proper scoring rule $\mathfrak{S} : \mathcal{R} \times \mathcal{O} \to \mathbb{R}$ and convex function $F : \mathcal{P} \to \mathbb{R}$ be given. Then there exists a proper scoring rule $\mathfrak{S}^{\mathcal{P}} : \mathcal{P} \times \mathcal{O} \to \overline{\mathbb{R}}$ and a function $\hat{p} : \mathsf{relint}(\mathcal{R}) \to \mathcal{P}$ such that*

$$\Gamma \circ \hat{p} \equiv \mathrm{id}, \quad \mathsf{Risk}(\mathfrak{S}^{\mathcal{P}}) = F, \text{ and } \mathfrak{S}^{\mathcal{P}}(\hat{p}(\cdot), \cdot) \equiv_{(\mathsf{relint}(\mathcal{R}), \mathcal{P})} \mathfrak{S}(\cdot, \cdot) \tag{4.24}$$

if and only if F is Γ-bounded and $\mathsf{Risk}(\mathfrak{S})$ is F-regular on $\mathsf{relint}(\mathcal{R})$.[8]

[8]Here implicitly $\Gamma(p) = p$ for the Bayes risk computation.

Proof. For shorthand, let $G \doteq \mathsf{Risk}(\mathfrak{S})$. Fixing $r \in \mathsf{relint}(\mathcal{R})$, Theorem 4.1 and fairness imply $\mathfrak{S}(r,p) = G(r) + \langle \Gamma(p) - r, dG_r \rangle$, so in particular $\mathfrak{S}(r,p) = G(r)$ for all $p \in \Gamma_r$.

For the "only if" direction let $\hat{p}$ and $\mathfrak{S}^{\mathcal{P}}$ be given satisfying condition (4.24), and note that we must have $F(p) = \mathfrak{S}^{\mathcal{P}}(p,p)$. Then we have

$$\forall r \in \mathsf{relint}(\mathcal{R}), \quad F(\hat{p}(r)) = \mathfrak{S}^{\mathcal{P}}(\hat{p}(r), \hat{p}(r)) = \mathfrak{S}(r, \hat{p}(r)) = G(r).$$

Now fix $r \in \mathsf{relint}(\mathcal{R})$, $p \in \Gamma_r$. Again by fairness and by definition of $\hat{p}$, we have

$$F(p) = \mathfrak{S}^{\mathcal{P}}(p,p) \geq \mathfrak{S}^{\mathcal{P}}(\hat{p}(r), p) = \mathfrak{S}(r,p) = G(r) = F(\hat{p}(r)),$$

which implies $G(r) = \inf_{p \in \Gamma_r} F(p)$ on $\mathsf{relint}(\mathcal{R})$ as desired. Finally, since $\hat{p}$ achieves the infimum, F must be Γ-bounded as well.

For the "if" direction, as F is Γ-bounded, for all $r \in \mathsf{relint}(\mathcal{R})$ we may take $\hat{p}(r) \in \operatorname{arginf}_{p \in \Gamma_r} F(p)$; this choice trivially satisfies $\Gamma \circ \hat{p} \equiv \mathrm{id}$. Since G is F-regular on $\mathsf{relint}(\mathcal{R})$, we must in particular have $G(r) = F(\hat{p}(r))$. By Theorem 2.1, we may take $\mathfrak{S}^{\mathcal{P}}(p,p) = F(p)$, which fully specifies $\mathfrak{S}^{\mathcal{P}}$ up to a choice of subgradient dF. To do this, for each $r \in \mathsf{relint}(\mathcal{R})$ we take $dF_{\hat{p}(r)} = \phi^\top dG_r$ (the other subgradients may be chosen arbitrarily), though we must first show that this is indeed a subgradient at $\hat{p}(r)$.

By Lemma 4.7, we have $G^*(d) = F^*(\phi^\top d)$. Also, $\langle r, d \rangle - G(r) = \left\langle \hat{p}(r), \phi^\top d \right\rangle - F(\hat{p}(r))$. Hence, applying Lemma 3.11 for G and F^*, we have

$$\begin{aligned} d \in \partial G(r) &\iff G^*(d) = \langle r, d \rangle - G(r) \\ &\iff F^*(\phi^\top d) = \left\langle \hat{p}(r), \phi^\top d \right\rangle - F(\hat{p}(r)) \\ &\iff \phi^\top d \in \partial F(\hat{p}(r)), \end{aligned}$$

so in particular, $dF_{\hat{p}(r)} \in \phi^\top dG_r \in \partial F(\hat{p}(r))$. Finally, we show that $\hat{p}$ does in fact induce $\mathfrak{S}$ from $\mathfrak{S}^{\mathcal{P}}$ for all $r \in \mathsf{relint}(\mathcal{R})$:

$$\begin{aligned} \mathfrak{S}^{\mathcal{P}}(\hat{p}(r), o) &= F(\hat{p}(r)) + dF_{\hat{p}(r)}(o) - dF_{\hat{p}(r)}(\hat{p}(r)) \\ &= G(r) + \langle \phi(o), dG_r \rangle - \langle \Gamma(\hat{p}(r)), dG_r \rangle \\ &= \mathfrak{S}(r, o). \end{aligned}$$

□

Theorem 4.14 gives precise conditions under which a score may be induced from another. The condition (4.24) is just expressing that $\hat{p}$ induces $\mathfrak{S}$ from $\mathfrak{S}^{\mathcal{P}}$, with the additional assumption that $\hat{p}$ be *calibrated*, meaning $\hat{p}(r)$ must itself be consistent with report r, i.e. $\Gamma(\hat{p}(r)) = r$. It is an interesting question whether this calibration property is a consequence of inducing $\mathfrak{S}$, or whether there are choices $\hat{p}$ which are not consistent in this way.

Note that the fairness assumption above is merely for convenience and is essentially without loss of generality. We could just have easily asked when $\mathfrak{S}^{\mathcal{P}}(\hat{p}(\cdot), \cdot) \cong \mathfrak{S}$,[9] as the

[9]Recall that scores are equivalent if their difference is a function solely of the observed outcome, and not the report.

answer would be exactly the same, except that we would replace $\mathsf{Risk}(\mathfrak{S})$ with the Bayes risk of a fair score equivalent to $\mathfrak{S}$.

The proof of Theorem 4.14 reveals something surprising — $\hat{p}$ is a generalized exponential family! To see this, note that we took $\hat{p}(r)$ such that $\phi^\top dG_r \in \partial F(\hat{p}(r))$, but by Corollary 3.14, this implies $\hat{p}(r) \in \partial F^*(\phi^\top dG_r)$, so $\hat{p}$ is a mean parameterization of an F-GEF. Hence, under our calibration assumption, GEFs are the unique family of distributions which induce scoring rules for means from classical scoring rules.

Another way to look at this is that every scoring rule for a linear property can be thought of as a scoring rule for a F-GEF for some other entropy F. To see this, just take $F(p) = G(\Gamma(p))$, which is convex. Hence we have actually answered question 1 above: the designer loses no flexibility by inducing scores rather than "starting from scratch."

We now briefly address questions 2 and 3, by appealing to the "overlapping bijections" discussion at end of § 4.3.1. In particular, the answer to question 2 is "yes", as any G is F-regular for $F : p \mapsto G(\mathbb{E}_p[\phi])$, while the answer to 3 is more complicated. In general, there is no succinct characterization of the different $G = \mathsf{Risk}(\mathfrak{S})$ that are F-regular, except for specific cases (like "log-exp-concave" for negative Shannon entropy, and piecewise quadratic for $\|\cdot\|^2$). However, the many-many relationship between F and G from § 4.3.1 transfers to the same relationship between scoring rules which can be induced from classical scoring rules.

Finally, it may seem that this technique of taking $\hat{p}(r) \in \operatorname{arginf}\{F(p) : p \in \Gamma_r\}$ generalizes to potentially nonlinear properties — could this method of inducing scores provide an answer to the questions posed in § 3.4.3 and show how to elicit nonlinear properties? As it turns out, this does not work in general. As a counter-example, take $\mathcal{P} = \Delta_3$, and define $\Gamma(p) = (p - x_0) \cdot (u - x_0)/\|p - x_0\|$, where $x_0 = [-\epsilon, (1-\epsilon)/2, (1-\epsilon)/2]$ and $u = [1/3, 1/3, 1/3]$ is the uniform distribution. Now take $F(p) = \|p\|^2$. Then when $p = [0, 1/2, 1/2]$, it is easy to see that $F(\hat{p}(r)) + \nabla F(\hat{p}(r))(p - \hat{p}(r)) = \|p\|^2 - \|p - \hat{p}(r)\|^2$ is locally *increasing* in r when moving away from $\Gamma(p)$, since the error is measured by Euclidean distance, and the selections $\hat{p}$ are moving closer to p — indeed, as we see in Figure 4.4, they almost reach p when ϵ is small.

Inducing prediction markets

We now turn to prediction markets, and ask similar questions here:

1. Can one run an incomplete market mechanism $\mathfrak{P}$ using a complete market $\mathfrak{P}^{\mathcal{P}}$?
2. Does the market designer lose flexibility by doing so?
3. In such a scheme, what will be the price space of $\mathfrak{P}^{\mathcal{P}}$?

Not surprisingly, in light of Theorem 4.3, our answers are very similar. First, we answer (1) in the affirmative, which as in the scoring rules setting, also answers (2) in the negative.

Theorem 4.15. *Let $C : \mathcal{V}^* \to \overline{\mathbb{R}}$ and $B : \mathcal{W}^* \to \overline{\mathbb{R}}$ be given convex, l.s.c., and proper functions, and define $\mathfrak{P}^{\mathcal{P}}(w^*, o) \doteq w^*(o) - B(w^*)$ and $\mathfrak{P}(v^*, o) \doteq \langle \phi(o), v^* \rangle - C(v^*)$. Then*

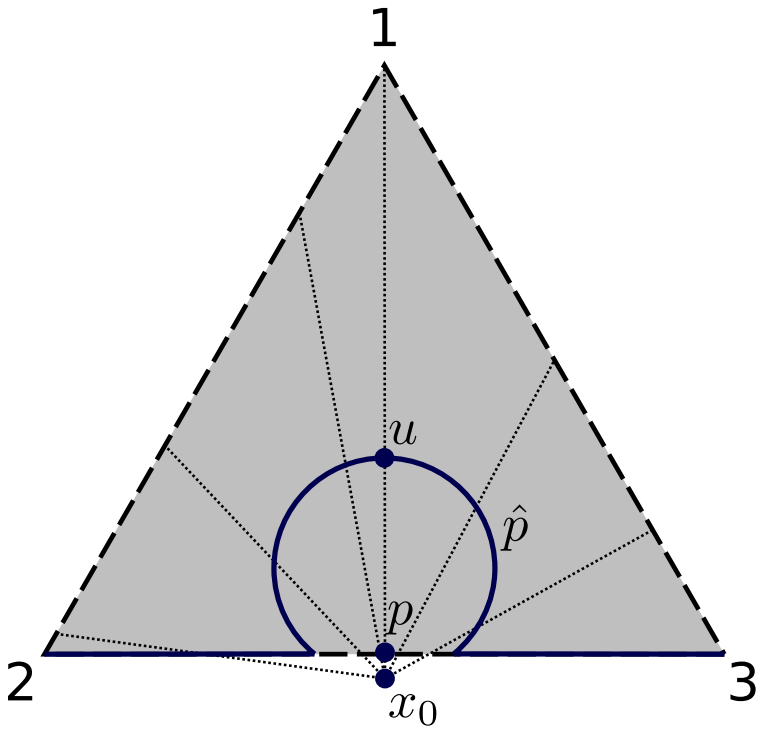


Figure 4.4: A 3-outcome example of a nonlinear Γ for which the maximum entropy $\hat{p}$ fails to induce a proper scoring rule. Recall any point on the circle with diameter $\overline{u\,x_0}$ forms a right angle with u and x_0, meaning it is the closest point to u on the ray out of x_0; this explains the peculiar form of $\hat{p}$ shown.

there is some $\hat{q} : \mathcal{V}^* \to \mathcal{W}^*$ *such that* $\mathfrak{P}(\cdot,\cdot) \equiv \mathfrak{P}^{\mathcal{P}}(\hat{q}(\cdot),\cdot)$ *if and only if* C^* *is* B^**-regular for statistic* ϕ*.*

Proof. We observe that the condition $\mathfrak{P}(\cdot,\cdot) \equiv \mathfrak{P}^{\mathcal{P}}(\hat{q}(\cdot),\cdot)$ breaks into two: (a) $\forall o \in \mathcal{O}$ $\hat{q}(v^*)(o) = \langle \phi(o), v^* \rangle$, and (b) $C(v^*) = B(\hat{q}(v^*))$. Condition (a) is equivalent to $\hat{q}(v^*) = \phi^\top v^*$ by definition. Using this, we can reduce (b) to $C(v^*) = B(\phi^\top v^*)$. Now letting $\theta \doteq v^*$, we can apply Lemma 4.7, and as has been argued previously (since the Fenchel–Moreau theorem holds under our assumptions), we see that (b) is equivalent to C^* being B^*-regular. □

While the proof of Theorem 4.15 does not mention GEFs explicitly, we can tease them out just as we did for Theorem 4.14. The trick lies in the answer to question (3): what is the price space for $\mathfrak{P}^{\mathcal{P}}$? We define this as ∂B, but we are interested in the possible prices attained by varying v^*, namely $\bigcup_{v^* \in \mathcal{V}^*} \partial B(\hat{q}(v^*)) = \bigcup_{\theta \in \mathcal{V}^*} \partial F^*(\phi^\top \theta)$, where $F = B^*$. Hence, if we pick a price p_θ corresponding to each θ, the prices form an F-GEF! In particular, if F^* is differentiable, the prices will be unique (often a desirable property for a prediction market mechanism in its own right), and will correspond to a unique F-GEF.

This latter observation deserves attention. When an incomplete prediction market $\mathfrak{P}$ is induced from a complete one $\mathfrak{P}^{\mathcal{P}}$, the prices are constrained to be a generalized exponential family, where the statistic ϕ is determined by, and indeed the same as, the payoff function of market $\mathfrak{P}$. Hence, traders with beliefs about the mean of the payoff function ϕ place bets in the market in exactly the same way as if they were betting directly on the outcomes, but with the prices constrained to a particular GEF. In particular, when $\mathfrak{P}^{\mathcal{P}}$ is LMSR, the most widely-used automated prediction market mechanism (see (1.5) and §1.2.2), agents are essentially trading on the mean parameters of classical exponential families! Furthermore, any prediction market with $C = G^*$ for a $(-H)$-regular G (or a G which is regular in the sense of Banerjee et al. [16]) can be expressed as an instance of LMSR in this way.

4.4 Discussion

We have just completed a winding tour through four seemingly unrelated concepts, and shown strong connections among them. To summarize what we have found, we give in Table 4.1 an overview of the rough conceptual or semantic relationships behind the bijections of Figure 4.1. It is especially interesting to correlate these relationships (and assumptions) with Figure 4.1, and in particular when a bijection crosses the duality line and when it does not.

Many observations can be made from this new vantage point. For example, it is clear from the assumptions needed (noting also that several are necessary in some respect) that the connection between scoring rules and divergences is the most fundamental; it is safe to say that these are essentially different representations of the same object.

More abstractly, though, it is clear that the driver behind nearly all of these results, and especially the more surprising ones, is convex duality. The functions G and C represent two sides of the same coin, but often our intuition is hesitant to flip from one side to another. This is perhaps especially true of the relationship between prediction markets and generalized exponential families — why should the prices be so structured when using a complete market to emulate an incomplete one?

Rather than enumerate further observations, we conclude our tour with another tour, to illustrate our newfound connections. Given an incomplete prediction market $\mathfrak{P}$ with payoff function ϕ, we have found that we can induce $\mathfrak{P}$ from a complete market $\mathfrak{P}^{\mathcal{P}}$ with cost function B exactly when the prices in $\mathfrak{P}^{\mathcal{P}}$ form a B^*-GEF P_Θ for *statistic* ϕ. We may move from $\mathfrak{P}$ to a scoring rule $\mathfrak{S}$ which is mutually reducible to $\mathfrak{P}$, which happens to be proper for the property $\Gamma(p) = \mathbb{E}_p[\phi]$, the mean of random variable ϕ. Asking the same question of induction for $\mathfrak{S}$ gives rise to the mean parameterization of the very same GEF P_Θ.

If we wish to measure the "distance" between beliefs or prices p in $\mathfrak{S}^{\mathcal{P}}$ or $\mathfrak{P}^{\mathcal{P}}$, using generalized relative entropy, or the "regret" faced by agents with one belief who act as if they had another, the answer can be expressed as a generalized Bregman divergence between the parameters, which are the reports and prices of $\mathfrak{S}$ and $\mathfrak{P}$, respectively. Finally, to come full

Bijection	Semantic relationship	Assumptions
DIV $\leftrightarrow$ SR Thm 4.1	A SR can be written as a DIV and vice versa.	(none)
PM $\leftrightarrow$ SR Cor 4.5	Trades in a PM yield the same payoff as a SR of the corresponding prices.	$\nabla G_{\mathrm{ri}(\mathcal{R})}$, G^{**}
PM $\leftrightarrow$ DIV Cor 4.6	Trades in a PM have net expected payoff equal to a divergence of their corresponding prices.	$\nabla G_{\mathrm{ri}(\mathcal{R})}$, G^{**}
DIV $\leftrightarrow$ GEF Thm 4.9	The generalized relative entropy of two members of a GEF is a DIV of their corresponding parameters.	$\langle\cdot,\cdot\rangle$, F^{**}, G^{**}, (G str cvx)
SR $\leftrightarrow$ GEF Thm 4.14	Any SR for the mean of a statistic can be written as a full-distribution SR applied to GEFs.	$\langle\cdot,\cdot\rangle$, F^{**}, G^{**}, F Γ-bdd
PM $\leftrightarrow$ GEF Thm 4.15	Payoffs in an incomplete market are identical to a complete market whose prices are restricted to a GEF.	F^{**}, G^{**}, $\langle\cdot,\cdot\rangle$

Table 4.1: The conceptual relationships between the four constructs presented in this chapter: generalized Bregman divergences (DIV), scoring rules (SR), prediction markets (PM), and generalized exponential families (GEF). All results assume that $\mathcal{P}$ is convex. For the other assumptions: $\nabla_{\mathrm{ri}(\mathcal{R})}G$ denotes the condition presented in Corollary 4.5; f^{**} denotes the requirement that f be proper and l.s.c.; $\langle\cdot,\cdot\rangle$ means the result requires a dual pair; and F Γ-bdd refers to the Γ-boundedness assumption, that F attains its lower bound on Γ_r for all $r \in \mathsf{relint}(\mathcal{R})$. For DIV $\leftrightarrow$ GEF, only the mean parameterization requires G to be strictly convex.

circle, we may in turn take these simple divergences and reconstruct the original prediction market $\mathfrak{P}$, scoring rule $\mathfrak{S}$, and B^*-GEF P_Θ.

We close with a look toward the future. Though comprehensive in a sense, our results beg the question particularly of generalized exponential families, as to how much of the extensive body of literature for classical families might extend to the general setting. Such extensions would immediately apply to the other three concepts discussed here. For example, we conjecture that the right notion of "generalized Bayesian updating" may yield a version of Vovk's aggregating algorithm for other entropies, perhaps shedding light on the theory of mixability in online learning.

Chapter 5

A new view of mechanism design

In this chapter, we apply our characterization results to mechanism design. As we will see, by viewing mechanisms in terms of convex analysis, we are able to make new insights, which either simplify or clarify existing results, or add new techniques.

We first examine a number of characterizations from the mechanism design literature that generally focus on when there exist payments that make a given allocation rule truthful. Figure 5.1 (a) illustrates these characterizations and how they were proved. As it shows, several of them rely on showing equivalence to a condition known as *cyclic monotonicity*. Instead, we translate these results into convex analysis terms and prove them by showing equivalence to the condition of being a family of subgradients of a convex function; see Figure 5.1 (b). This has two main benefits. First, since cyclic monotonicity is a difficult condition to work with, we are able to greatly simplify the proofs of these results. Second, our proofs generally proceed by explicitly constructing the convex function, which gives a natural characterization of the payments rather than just showing that they exist. This approach also illuminates how a result by Carroll [28] about truthful mechanisms is essentially a similar characterization (see Theorem 5.3).

We then combine these results with our results about properties from Chapter 3, to obtain new truthfulness checks for the case where there are finitely many outcomes. Finally, we conclude with remarks about revenue equivalence and future work.

5.1 Implementability conditions via convexity

Interpreted in the general mechanism design framework given by Definition 2.3, Theorem 2.1 says that a mechanism (f, p) is truthful if and only if the consumer surplus function $t \mapsto U(t, t)$ that it implicitly defines on the convex hull of $\mathcal{T}$ is a convex function which has a subgradient consistent with f on $\mathcal{T}$. This is a known characterization for the case of convex $\mathcal{T}$ (as well as non-convex $\mathcal{T}$ that satisfy an assumption known as outcome compactness) [7], but in practice the consumer surplus function is not always the most natural representation of a mechanism. In this section, we examine two other approaches to characterizing truthful

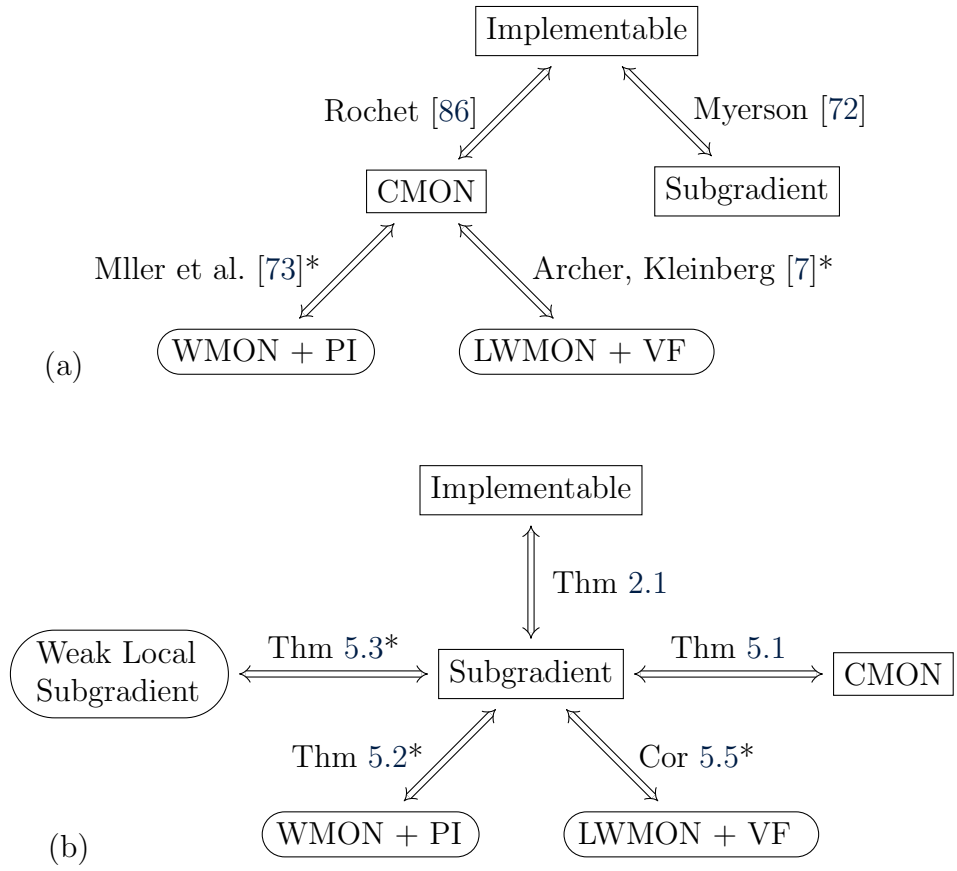


Figure 5.1: Proof structure of existing mechanism design literature (a), and the new proof structure presented in this dissertation (b). Rounded rectangles and asterisks denote the requirement that $\mathcal{T}$ be convex.

mechanisms that have been explored in the literature and show that they have insightful interpretations in convex analysis. This interpretation has two benefits. First, by focusing on the essential convex analysis questions we are able to greatly simplify many of the proofs. Second, our proofs are constructive; in many cases we explicitly construct a consumer surplus function G, which when the mechanism is being represented by its allocation rule gives the necessary payments rather than simply providing a proof that payments exist.

5.1.1 Subgradient characterizations

From an algorithmic perspective, it may be more natural to focus on the design of the allocation rule f rather than the specific payments. There is a large literature that focuses on when there exists a choice of payments p to make f into a truthful mechanism (e.g. [91, 8]). Since such payments exist if and only if there is a convex function for which f is a

subgradient at points in $\mathcal{T}$, this is essentially a very natural convex analysis question: when is a function f a subgradient of a convex function? Unsurprisingly, the central result in the literature is closely connected to convex analysis.

Definition 5.1. *A family* $\{dG_t \in \mathsf{Lin}(\mathcal{V} \to \mathbb{R})\}_{t\in\mathcal{T}}$ *satisfies* cyclic monotonicity (CMON) *if for all finite sets* $\{t_0, \dots, t_k\} \subseteq \mathcal{T}$*,*

$$\sum_{i=0}^{k} dG_{t_i}(t_{i+1} - t_i) \leq 0, \tag{5.1}$$

where indices are taken modulo $k+1$*. We refer to the weaker condition that* (5.1) *hold for all pairs* $\{t_0, t_1\}$ *as* weak monotonicity (WMON)*.*

A well known characterization from convex analysis is that a function f defined on a convex set is a subgradient of a convex function on that set iff it satisfies CMON [89]. Rochet's [86] proof that such payments exist on a possibly non-convex $\mathcal{T}$ iff f satisfies CMON is effectively a proof of the following generalization of this theorem.

Theorem 5.1. *A family* $\{dG_t \in \mathsf{Lin}(\mathcal{V} \to \mathbb{R})\}_{t\in\mathcal{T}}$ *satisfies CMON if and only if there exists a convex* $G : \mathsf{conv}(\mathcal{T}) \to \mathbb{R}$ *such that* dG_t *is a subgradient of* G *at* t *for all* $t \in \mathcal{T}$*.*

Rochet notes that his proof is adapted from the one given in Rockafellar's text [89] of the weaker theorem where $\mathcal{T}$ is restricted to be convex. We adapt Rochet's proof to highlight how its core is a construction of G. .

Proof. Given such a G, by (2.2) we have $dG_{t_i}(t_{i+1} - t_i) \leq G(t_{i+1}) - G(t_i)$. Summing gives (5.1). Given such a family $\{dG_t\}_{t\in\mathcal{T}}$, fix some $t_0 \in \mathcal{T}$ and define

$$G(t) = \sup_{\substack{\{t_1,\dots,t_{k+1}\}\subseteq\mathcal{T},\\ t_{k+1}=t}} \sum_{i=0}^{k} dG_{t_i}(t_{i+1} - t_i), \tag{5.2}$$

where $\{t_1, \dots, t_k\}$ denotes any finite sequence (k is not fixed).

By CMON, for $t \in \mathcal{T}$ this sum is upper bounded by $-dG_t(t_0 - t)$. Thus, the supremum is finite on $\mathcal{T}$. G is a pointwise supremum of convex functions, so is convex. By convexity, G is also finite on $\mathsf{conv}(\mathcal{T})$.

For any $t \in \mathcal{T}$ and $t' \in \mathsf{conv}(\mathcal{T})$,

$$\begin{aligned} G(t) + dG_t(t' - t) &= dG_t(t' - t) + \sup_{\substack{\{t_1,\ldots,t_{k+1}\} \subseteq \mathcal{T},\\ t_{k+1}=t}} \sum_{i=0}^{k} dG_{t_i}(t_{i+1} - t_i) \\ &= \sup_{\substack{\{t_1,\ldots,t_{k+1}\} \subseteq \mathcal{T},\\ t_k=t,\, t_{k+1}=t'}} \sum_{i=0}^{k} dG_{t_i}(t_{i+1} - t_i) \\ &\leq \sup_{\substack{\{t_1,\ldots,t_{k+1}\} \subseteq \mathcal{T},\\ t_{k+1}=t'}} \sum_{i=0}^{k} dG_{t_i}(t_{i+1} - t_i) \\ &= G(t'), \end{aligned}$$

so dG_t satisfies (2.2). □

A number of papers have sought simpler and more natural conditions than CMON that are necessary and sufficient in special cases, e.g. [91, 7, 8]. These results are typically proven by showing they are equivalent to CMON. However, it is much more natural to directly construct the relevant G. This also often has the advantage of providing a characterization of the payments that is more intuitive than the supremum in Rochet's construction. As an example, we show one such result has a simple proof using our framework.

As in Myerson's [72] construction for the single-parameter case, we construct a G by integrating over dG_t. In particular, for any two types x and y our construction makes use of the line integral

$$\int_{L_{xy}} dG_t(y - x)dt = \int_0^1 dG_{(1-t)x+ty}(y - x)dt.$$

As Berger et al. [20] and Ashlagi et al. [8] observed, if $\{dG_t\}_{t\in\mathcal{T}}$ satisfies WMON and $\mathcal{T}$ is convex, this (Riemann) integral is well defined because it is the integral of a monotone function. If these line integrals vanish around all triangles (equivalently $\int_{L_{xy}} dG_t(y - x)dt + \int_{L_{yz}} dG_t(z - y)dt = \int_{L_{xz}} dG_t(z - x)dt$) we say $\{dG_t\}$ satisfies *path independence*.

Theorem 5.2 (adapted from [73])**.** *For convex $\mathcal{T}$, a family $\{dG_t \in \mathsf{Lin}(\mathcal{V} \to \mathbb{R})\}_{t\in\mathcal{T}}$ is a subgradient of a convex function if and only if $\{dG_t\}_{t\in\mathcal{T}}$ satisfies WMON and path independence.*

Proof. Given a convex function G and $\{dG_t\}$, $\{dG_t\}$ satisfies CMON and thus WMON. Path independence also follows from convexity (Rockafellar [89] p. 232). Now given a $\{dG_t\}$ that satisfies WMON and path independence, fix a type $t_0 \in \mathcal{T}$ and define $G(t') = \int_{L_{t_0t'}} dG_t(t' - t_0)dt$ (well defined by WMON as the integral of a monotone function). Given $x, y, z \in \mathcal{T}$ such that $z = \lambda x + (1 - \lambda)y$, by path independence and the linearity of dG_z we

have

$$\begin{aligned}
&\lambda G(x) + (1-\lambda)G(y) \\
&= G(z) + \lambda \int_{L_{zx}} dG_t(x-z)dt + (1-\lambda)\int_{L_{zy}} dG_t(y-z)dt \\
&\geq G(z) + \lambda dG_z(x-z) + (1-\lambda)dG_z(y-z) = G(z),
\end{aligned}$$

so G is convex. Similarly, for $x, y \in \mathcal{T}$, dG_t satisfies (2.2) because

$$dG_x(y-x) \leq \int_{L_{xy}} dG_t(y-x)dt = G(y) - G(x). \qquad \square$$

5.1.2 Local conditions

In many settings, it is natural to specify mechanisms in terms of an algorithm that computes the allocation and payment. As it is often easier to reason about the behavior of algorithms given small changes to their input rather than arbitrary changes, several authors have sought to characterize truthful mechanisms using local conditions [7, 20, 28].

We show in this section how many of these results are in essence a consequence of a more fundamental statement, that convexity is an inherently local property. For example, in the twice differentiable case it can be verified by determining whether the Hessian is positive semidefinite at each point. We start with a local convexity result, and use it to show that an affine score is truthful if and only if it satisfies a very weak local truthfulness property introduced by Carroll [28]. Afterwards we turn to a similar characterization by Archer and Kleinberg [7].

Definition 5.2. *Given a function G, a family $\{dG_t \in \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}})\}_{t\in\mathcal{T}}$ is a* weak local subgradient with respect to G (G-WLSG) *if for all $t \in \mathcal{T}$ there exists an open neighborhood U_t of t such that for all $t' \in U_t$,*

$$G(t) \geq G(t') + dG_{t'}(t-t') \quad \text{and} \quad G(t') \geq G(t) + dG_t(t'-t). \tag{5.3}$$

We now show that G-WLSG is a sufficient condition for a family of functions to be a subgradient of G. The proof is heavily inspired by Carroll [28].

Theorem 5.3. *Let $\mathcal{T}$ be convex. A family $\{dG_t \in \mathsf{Lin}(\mathcal{V} \to \overline{\mathbb{R}}\}_{t\in\mathcal{T}}$ is a subgradient of a given function G if and only if it satisfies G-WLSG.*

(Adapted from [28]). As usual, the forward direction is trivial. For the other, let $t, t' \in \mathcal{T}$ be given; we show that the subgradient inequality for $dG_{t'}$ holds at t. By compactness of $\mathsf{conv}(\{t, t'\})$, we have a finite set $t_i = \alpha_i t' + (1-\alpha_i)t$, where $0 = \alpha_0 \leq \cdots \leq \alpha_{k+1} = 1$, such that G-WLSG holds between each t_i and t_{i+1}. (The cover $\{U_s \mid s \in \mathsf{conv}(\{t, t'\})$ has a finite

subcover. Take t_{2i} from the subcover and $t_{2i+1} \in U_{t_{2i}} \cap U_{t_{2i+2}}$.) By the WLSG condition (5.3), we have for each i,

$$0 \geq G(t_{i+1}) - G(t_i) + dG_{t_{i+1}}(t_i - t_{i+1}) \tag{5.4}$$
$$0 \geq G(t_i) - G(t_{i+1}) + dG_{t_i}(t_{i+1} - t_i). \tag{5.5}$$

Now using the identity $t_{i+1} - t_i = (\alpha_{i+1} - \alpha_i)(t' - t)$ and adding $\alpha_i/(\alpha_{i+1} - \alpha_i)$ times (5.4) to $\alpha_{i+1}/(\alpha_{i+1} - \alpha_i)$ times (5.5), we have

$$0 \geq G(t_i) - G(t_{i+1}) + \alpha_i dG_{t_i}(t' - t) - \alpha_{i+1} dG_{t_{i+1}}(t' - t). \tag{5.6}$$

Summing (5.6) over $0 \leq i \leq k$ gives

$$0 \geq G(t_0) - G(t_{k+1}) + \alpha_0 dG_{t_0}(t' - t) - \alpha_{k+1} dG_{t_{k+1}}(t' - t),$$

which when recalling our definitions for α_i and t_i yields the result. □

The WLSG condition translates to an analogous notion in terms of truthfulness, *weak local truthfulness*.

Definition 5.3. *An affine score is* weakly locally truthful *if for all $t \in \mathcal{T}$ there exists some open neighborhood U_t of t, such that truthfulness holds between t and every $t' \in U_t$, and vice versa. That is,*

$$\forall t \in \mathcal{T},\ \forall t' \in U_t,\ \ S(t')(t) \leq S(t)(t) \ \ \textit{and} \ \ S(t)(t') \leq S(t')(t'). \tag{5.7}$$

Corollary 5.4 (Generalization of Carroll [28])**.** *An affine score $S : \mathcal{T} \to \mathcal{A}$ for convex $\mathcal{T}$ is truthful if and only if it is weakly locally truthful.*

Proof. Defining $G(t) \doteq S(t)(t)$, by weak local truthfulness we may write

$$G(t) = S(t)(t) \geq S(t')(t) = G(t') + S_\ell(t')(t - t')$$
$$G(t') = S(t')(t') \geq S(t)(t') = G(t) + S_\ell(t)(t' - t),$$

where t' is local to t and $S_\ell(\cdot)$ is the linear part of $S(\cdot)$. This says that $dG_t = S_\ell(t)$ satisfies G-WLSG; the rest follows from Theorem 5.3 and Theorem 2.1. □

Finally, in the spirit of Section 5.1.1, Archer and Kleinberg [7] characterized local conditions under which an allocation rule can be made truthful. A key condition from their paper is *vortex-freeness*, which is a condition they show to be equivalent to local path independence. The other condition, local WMON, means that WMON holds in some neighborhood around each type. Their result follows directly from an analogous characterization of subgradients.

Corollary 5.5. *Let $\mathcal{T}$ be convex. A family $\{dG_t \in \mathsf{Lin}(\mathcal{V} \to \mathbb{R})\}_{t \in \mathcal{T}}$ is a subgradient of a convex function if and only if it satisfies local WMON and is vortex-free.*

Proof. We prove the reverse direction; suppose $\{dG_t\}_{t\in\mathcal{T}}$ satisfies local WMON and is vortex-free. From Lemma 3.5 of [7] we have that vortex-freeness is equivalent to path independence, so by Theorem 5.2 for all t there exists some open U_t such that $\{dG_{t'}\}_{t'\in U_t}$ is the subgradient of some convex function $G^{(t)} : U_t \to \mathbb{R}$. We need only show the existence of some G such that $\{dG_t\}_{t\in\mathcal{T}}$ is the subgradient of G on each U_t; the rest follows from Theorem 5.3.

Fix some $t_0 \in \mathcal{T}$ and define $G(t) = \int_{L_{t_0\,t}} dG_{t'}dt'$, which is well defined by compactness of $\mathsf{conv}(\{t_0, t\})$ and the fact that a locally increasing real-valued function is increasing. But for each t' and $t \in U_{t'}$ we can also write $G^{(t')}(t) = \int_{L_{t'\,t}} dG_{t''}dt''$ by [89, p. 232], and now by path independence we see that G and $G^{(t')}$ differ by a constant. Hence $\{dG_t\}_{t\in\mathcal{T}}$ must be a subgradient of G on $U_{t'}$ as well, for all $t' \in \mathcal{T}$. □

5.2 Properties in mechanism design

In the case when R is finite, we can combine Theorem 3.17 and Corollary 5.4 to obtain the following simple truthfulness check. This essentially says that one need only check overlapping report regions for r and r' (i.e. r and r' are both correct reports for some type t) to make sure that any type with r correct and r' incorrect yields a lower utility for r' than r.

Theorem 5.6. *For finite R, an affine score $S : R \to \mathcal{A}$ elicits Γ if and only if for all r, r' such that $\Gamma_r \cap \Gamma_{r'} \neq \emptyset$, we have $S(r')(t) < S(r)(t)$ for all $t \in \Gamma_r \setminus \Gamma_{r'}$.*

We now apply our results to mechanism design with a finite set of allocations. One example of such a setting is ordinal utilities over a finite set of outcomes. Let $\mathcal{T} = (\mathcal{O} \to \mathbb{R})$ and let R be the set of preference orderings over $\mathcal{O}$, which we represent as permutations π of $\mathcal{O}$ (higher preference being first). Then a natural property is

$$\Gamma(t) = \{\pi \in R \mid \forall i < j,\ t(\pi_i) \leq t(\pi_j)\}, \tag{5.8}$$

the set of rankings consistent with type t. Note that unless there are ties, meaning $t(o) = t(o')$ for some $o, o' \in \mathcal{O}$, this mapping will be a singleton. More generally, truncated rankings or any mechanism with a finite message space yields such a property.

Carroll [28] points out that it is known that for all finite sets of allocations, the set of types for which a particular allocation is optimal forms a polyhedron and then shows that for all convex set of types it is sufficient to check incentive compatibility constraints between polyhedra that intersect at a face. Applying Theorem 3.17 shows something stronger: the types form not just an arbitrary set of polyhedra, but a power diagram. This observation yields the following non-implementability check.

$$\text{Power diagram test:} \quad \text{The sets } \text{cell}_o := \{t | f(t) = o\} \text{ form a power diagram.} \tag{5.9}$$

Thus, if (5.9) fails, f is not implementable.

To illustrate this test, consider the following two allocation functions.

$$f^1(t) = \begin{cases} \emptyset & \text{if } t_a < 2 \text{ and } t_b < 2 \\ \{a\} & \text{if } t_a \geq 2 \text{ and } t_b < 2 \\ \{b\} & \text{if } t_a < 2 \text{ and } t_b \geq 2 \\ \{a, b\} & \text{if } t_a \geq 2 \text{ and } t_b \geq 2 \end{cases} \tag{5.10}$$

$$f^2(t) = \begin{cases} \emptyset & \text{if } t_a < 2 \\ \{a\} & \text{if } t_a \geq 2 \text{ and } t_b < 2 \\ \{a, b\} & \text{if } t_a \geq 2 \text{ and } t_b \geq 2 \end{cases} \tag{5.11}$$

As Figure 5.2 shows, f^1 satisfies (5.9), as it partitions the type space $\mathcal{T} = \mathbb{R}^2$ into a power diagram; just take sites $P = \{(1,1),(1,3),(3,1),(3,3)\}$ and weights $w_i = -\|p_i\|/2$. Moreover, f^1 is implementable. The allocation function f^2, however, fails the test: no finite site exists for the cell $\{t|t_a < 2\}$, since adjacent cell boundaries are perpendicular to the line between their sites. Hence we can immediately conclude that f^2 is not implementable, and in fact *any* allocation function which partitions the type space in the same way (i.e. uses the same cases in (5.11) with different values) would fail to be implementable.

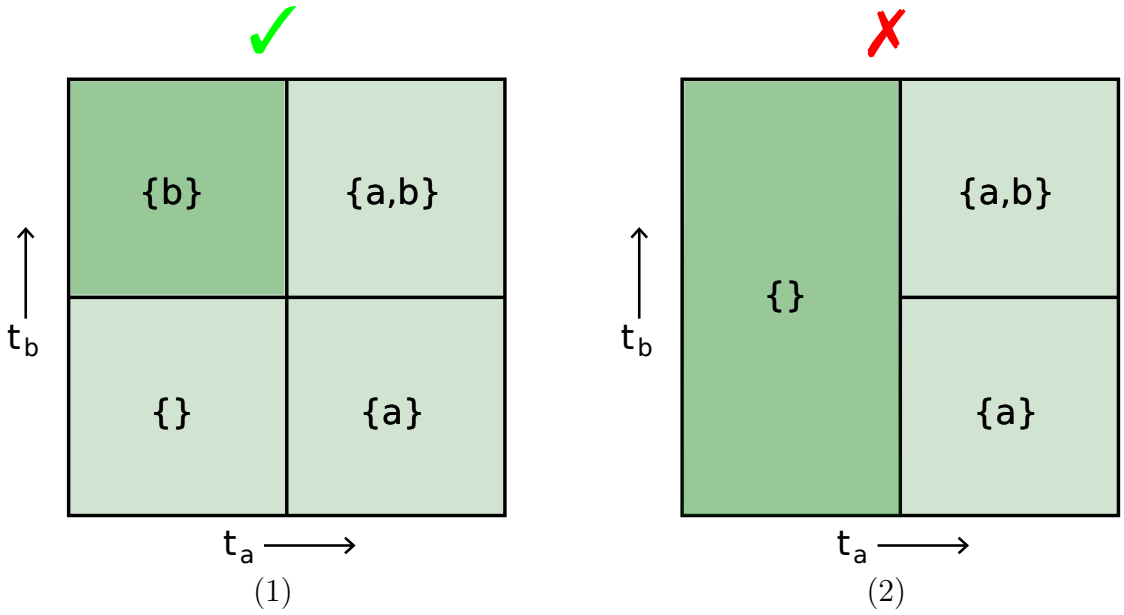


Figure 5.2: A mechanism that passes the power diagram test (1) and one that fails (2).

We can now see why polyhedral type spaces are too weak: the allocation function f^2 is not implementable, yet clearly has a polyhedral type space. The power diagram condition (5.9), then, is a stronger necessary condition for truthfulness. Note however that it is not sufficient, as the power diagram test is not sensitive to the actual allocations chosen — just swap $\{a\}$

and $\{a, b\}$ for f^1, and the resulting allocation function still passes the test but is no longer implementable. In is true, though, that every power diagram is the type space for some implementable allocation function.

It remains to show how the test (5.9) can be implemented. In fact, we have already discussed this in §3.4.1; recall that for the *simple* case, Aurenhammer gives an algorithm to detect whether a set C of cells form a power diagram in §2.2 of [10], which runs in time $O(m)$, where m is the number of facets (cell faces of dimension $n-1$, where n is the number of outcomes). For the general case we may use the polynomial-time algorithm given by Rybnikov in [90, §12].

5.3 Revenue equivalence

Perhaps the most celebrated result in auction theory is the revenue equivalence theorem, which states that, in a single item auction, the revenue from an agent (equivalently that agent's consumer surplus) is determined up to a constant by the equilibrium probability that each possible type of that agent will receive the item [72]. A large body of work has looked for more general conditions under which this property holds (see, e.g., [61, 57]). Many of these conditions imply the convexity of the set of types. Our main result can be used to provide intuition for this.

Theorem 5.7. *Let $\mathcal{T}$ be convex, a truthful affine score $S : \mathcal{T} \to \mathcal{A}$ be given, and $\{dG_t\}_{t\in\mathcal{T}}$ be the corresponding selection of subgradients from (2.3). Then any truthful affine score $S' : \mathcal{T} \to \mathcal{A}$ with the same corresponding selection of subgradients differs from S by a constant (i.e. $S(t')(t) = S'(t')(t) + c$).*

Proof. By Theorem 2.1, we know that S and S' only differ only in their choice of convex function G. However, each choice has the same selection of subgradients, and two convex functions with the same selection of subgradents differ by a constant [89]. For intuition, see the construction of G by integrating its subgradients in the proof of Theorem 5.2. □

Note how the convexity of $\mathcal{T}$ is crucial here. Theorem 2.1 requires the existence a convex G on $\mathsf{conv}(\mathcal{T})$. If $\mathcal{T}$ is not convex, there may exist convex G and G' that share the same selection of subgradients on $\mathcal{T}$ but differ on $\mathsf{conv}(\mathcal{T}) - \mathcal{T}$, in which case they need not be revenue equivalent. The following example from [57] is illustrative.

There is an agent who has some demand d for a good. If he receives x units of the good his utility is $u_d(x) \doteq \min(0, x - d)$: he has a utility of 0 when he receives at least d units and $x - d < 0$ otherwise. Even if the set of possible values of d is a convex set, say $[0, 1]$, this type space $\mathcal{T} = \{u_d(x) | 0 \le d \le 1\}$ is not convex in our sense — for example, $(u_{0.3} + u_{0.4})/2 \notin \mathcal{T}$.

Consider a mechanism that gives the agent exactly his demand d, i.e. $f(u_d) = d$. One way to extend this onto the convex hull of the type space is to have the mechanism give the agent the minimum amount of the good this is needed for him to have a utility of 0, namely

$f^1(\sum \alpha_i u_{d_i}) = \max\{d_i\}$. Note that crucially $f^1|_{\mathcal{T}} = f$. With f^1, the payment $p^1(x)$ the agent makes is constant regardless of his allocation.[1]

An alternate way to extend this onto the convex hull is to set a schedule of prices such that an agent given x units of the good pays $p^2(x) = x + c$ and then give an agent of type $v = \sum \alpha_i u_{d_i}$ his optimal amount of the good (breaking ties in favor of giving more of the good): $f^2(v) = \max(\text{argmax}_x\{v(x) - (x + c)\})$. Despite once again having $f^2|\mathcal{T} = f$, the payment of an agent with type $v = u_d$ is now $p^2(d) = d+c$, which depends on the demand and hence is not $p^1(d) = c$. Thus, revenue equivalence does not hold for our original allocation rule f on $\mathcal{T}$.

In our example, there were several different ways to extend the allocation onto the convex hull of the type space such that the mechanism was implementable, so the mechanism did not satisfy revenue equivalence. Clearly convexity of $\mathcal{T}$ is a sufficient condition to ensure that there is a unique extension onto $\mathsf{conv}(\mathcal{T})$, but it is not a necessary one; Heydenreich et al. give several examples of conditions on allocation rules for this example that ensure they do satisfy revenue equivalence. Even having a unique extension is actually sufficient but not necessary. In particular, if some type $t \in \mathsf{conv}(\mathcal{T}) - \mathcal{T}$ is indifferent among several outcomes, it may be possible to change that type's allocation without changing the resulting G. In terms of convex analysis, this is because, under mild conditions, a convex function can be represented up to a constant by *any* selection of its subgradients (see [61]).

5.4 Future work

Below we detail several possible avenues for future work.

Multi-agent settings. The reader may note that thus far we have considered only single-agent mechanisms. As we argued in §2.3.1, this is in some sense without loss of generality, but in practice many constraints on mechanisms are crucially interdependent among agents; e.g. a single item which can only be given to one agent, or a combinatorial auction with nontrivial feasibility constraints. To have bite in such settings, we will have to extend our model and characterizations explicitly.

One way to extend the model is as follows. We have type spaces $\mathcal{T}_i$ for each agent, and joint type space $\mathcal{T} = \otimes\mathcal{T}_i$. The affine spaces are $\mathcal{A}_i \subseteq \mathsf{Aff}(\mathcal{T}_i \to \mathbb{R})$ with $\mathcal{A} \subseteq \otimes\mathcal{A}_i$. Finally, the utility of agent i with type t_i upon reports t' is $\mathfrak{A}(t')(t)_i = \mathfrak{A}(t')_i(t_i)$.

While it is common in mechanism design to think of a single outcome being chosen, thus dictating the utilities of all agents, here the mechanism would seem to have more flexibility, in choosing $a_i \in \mathcal{A}_i$ more or less independently of $j \neq i$. However, note that by being a subset of the cartesian product, $\mathcal{A}$ can encode constraints of the mechanism. In particular, $\mathcal{A}$ could be the set of utilities $\{a_{i,o}\}_{i\in[n]}$ corresponding to each outcome of some set $\mathcal{O}$, thus reducing to the common mechanism design case.

[1] An agent of type u_d can always report that he is of type $(1 - \epsilon)u_0 + \epsilon u_d$. This type has an arbitarily small value for receiving the good, but still receives d units according to f^1.

One of the most important results in mechanism design is Roberts' theorem, which says that for an unrestricted type space (i.e. $\mathcal{T} = \mathbb{R}^n_+$ for n agents), the only implementable allocation functions are *affine maximizers*, taking the following form:

$$f(t) \in \operatorname*{argmax}_{o \in \mathcal{O}} \sum_{i=1}^{n} w_i t_i(o) + C_o, \tag{5.12}$$

for some weights w_i and constants C_o. This is often viewed as an impossibility result, but it is known that as one restricts the type space, other possibilities arise in the form of f. Naturally, then, recent research has strived to characterize the form of implementable functions in restricted domains. As a first step in this vein, many new proofs of Roberts' theorem have appeared (see e.g. [68, 67, 41]) with an eye toward modularity, in the hopes that the right techniques would extend more easily into restricted domains. Our hope is that the general convexity-driven approach taken in this dissertation would help in this regard; we have already seen that our techniques greatly simplify the single-agent case, and given that a truthful multi-agent mechanism yields a truthful mechanism when fixing $n-1$ of the agents' types, it seems likely that these techniques could be used to simplify or extend existing multi-agent characterizations.

Beyond mechanism design, other multi-agent elicitation models have appeared in the literature, most notably "wagering mechanisms" from Lambert et al. [64], which in essence are multi-agent scoring rules. In the setting they analyze, the mechanism is constrained to ensure that for some budget B, the total payout is exactly B no matter what outcome ω materializes. One could view such a mechanism as an affine score, setting $\mathcal{A}_i \subseteq \{t_i \mapsto \mathbb{E}_{\omega \sim t_i}[s_i] \mid s_i \in \mathbb{R}^\Omega\}$. Note that this in some sense corresponds to a "transfer-free" mechanism, as $\mathcal{A}_i$ is actually a set of linear functions, so the affine term (the "payment") is 0. Such multi-agent scoring rules only become interesting when the mechanism has constraints among the agents; in this case, the budget constraint requires $\sum_i s_i = B$ for all outcomes $\omega \in \Omega$.

Envy-free mechanisms and other externalities. In several settings in mechanism design, one is conerned with agent utilities which depend not only on the chosen allocation, but possibly on the types of the other agents. For example, when participating in an auction for a nuclear weapon, my utility may depend on who gets the weapon, *and their value for it*. Note however that we have constructed $\mathcal{A}$ so that $\mathfrak{A}(t')(t)_i$ depends only on t' and t_i (the true type of agent i). Thus, to capture these interdependencies among types, we may simply lift our restriction on $\mathcal{A}$ by requiring only $\mathcal{A} \subseteq \mathsf{Aff}(\mathcal{T} \to \mathbb{R}^n)$. This would cover settings such as envy-free mechanisms, as well as other types of externalities, such as those in Fiat et al. [45]. Given how natural it is to express these settings in our model, it seems promising to apply our techniques here as well.

Revenue equivalence. Recently results have emerged characterizing revenue equivalence, specifically, on which domains it does and does not hold; see e.g. [57]. Given the discussion in

§5.3, it is likely that our approach would be more direct, yielding shorter or more constructive proofs.

Properties. Our exploration of properties in mechanism design has just scratched the surface. One promising direction is to draw connections to existing notions in mechanism design which seem to bear some resemblence to properties, such as multidimensional screening (see e.g. [87]). Another is to use the power diagram test (5.9) and other insights to derive new negative results for multi-agent settings — in fact, one could regard Hyafil and Boutilier [58] as a prelude to this general approach. Finally, Saks and Yu [91] proved that all monotone deterministic allocation rules are implementable when the type space is convex and the number of outcomes is finite; it may be that our approach using finite-valued properties will yield a simpler and more intuitive proof of this result.

Report duality in mechanism design. It is interesting to ask whether the notion of report duality, introduced in §3.3.2, can aid in the design of mechanisms. Such a technique could proceed by first choosing G^* and then taking $G = G^{**}$, just as is now common in the prediction markets literature. This approach may help meet certain constraints; if one wants any truthful mechanism with allocations in some set $\mathcal{O}$, one could choose a convex function G^* over $\mathsf{conv}(\{\mathsf{Eval}_o | o \in \mathcal{O}\})$ and work backwards to get the allocations by taking $f := dG^{**}$.

Chapter 6

Interpreting prediction markets

In this chapter, we strengthen recent connections between prediction markets and learning by showing that a natural class of market makers can be understood as performing stochastic mirror descent when trader demands are sequentially drawn from a fixed distribution. This provides new insights into how market prices (and price paths) may be interpreted as a summary of the market's belief distribution by relating them to the optimization problem being solved. In particular, we show that under certain conditions the stationary point of the stochastic process of prices generated by the market is equal to the market's Walrasian equilibrium of classic market analysis. Together, these results suggest how traditional market making mechanisms might be replaced with general purpose learning algorithms while still retaining guarantees about their behavior.

6.1 Introduction and literature review

This chapter is part of an ongoing line of research, spanning several authors, into formal connections between markets and machine learning. In Chen and Vaughan [34] an equivalence is shown between the theoretically popular prediction market makers based on sequences of proper scoring rules and follow the regularised leader, a form of no-regret online learning. By modelling the traders that demand the assets the market maker is offering we are able to extend the equivalence to stochastic mirror decent. The dynamics of wealth transfer is studied in Beygelzimer et al. [22], for a sequence of markets between agents that behave as Kelly bettors (i.e. have log utilities), and an equivalence to stochastic gradient decent is analysed. More broadly, Storkey [94] and Barbu and Lay [17] have analysed how a wide range of machine learning models can be implemented in terms of market equilibria.

The literature on the interpretation of prediction market prices, including Manski [70] and Wolfers and Zitzewitz [102] has had the goal of relating the equilibrium prices to the distribution of the beliefs of traders. More recent work of Othman and Sandholm [81] has looked at a stochastic model, and studied the behavior of simple agents sequentially interacting with the market. We continue this latter path of research, motivated by the

observation that the equilibrium price may be a poor predictor of the behavior in a volitile prediction market. As such, we seek a more detailed understanding of the market than the equilibrium point – we would like to know what the "stationary distribution" of the price is, as time goes to infinity.

As is standard in the literature, we assume a fixed (product) distribution over traders' beliefs and wealth. Our model features an automated market maker, following the framework of Abernethy et al. [1] is becoming a standard framework in the field.

We obtain two results. First, we prove that under certain conditions the stationary point of our stochastic process defined by the market maker and a belief distribution of traders converges to the Walrasian equilibrium of the market as the market liquidity increases. This result, stated in Theorem 6.1, is general in the sense that only technical convergence conditions are placed on the demand functions of the traders – as such, we believe it is a generalisation of the stochastic result of [81] to cases where agents are are not limited to linear demands, and leave this precise connection to future work.

Second, we show in Corollary 6.4 that when traders are Kelly bettors, the resulting stochastic market process is equivalent to stochastic mirror descent; see e.g. [42]. This result adds to the growing literature which relates prediction markets, and automated market makers in general, to online learning; see e.g. [1, 34, 22].

This connection to mirror descent seems to suggest that the prices in a prediction market at any given time may be meaningless, as the final point in stochastic mirror descent often has poor convergence guarantees. However, standard results suggest that a prudent way to form a "consensus estimate" from a prediction market is to *average* the prices. The average price, assuming our market model is reasonable, is provably close to the stationary price. In §6.5 we give a natural example that exhibits this behavior. Beyond this, however, Theorem 6.3 gives us insight into the relationship between the market liquidity and the convergence of prices; in particular it suggests that we should increase liquidity at a rate of $\sqrt{t}$ if we wish the price to settle down at the right rate.

6.2 Model

Our market model will follow the automated market maker framework of Abernethy et al. [1], as described in § 1.2.2, though our notation will differ slightly. We will equip our market maker with a strictly convex function $C : \mathbb{R}^n \to \mathbb{R}$ which is twice continuously differentiable. For brevity we will write $\varphi \doteq \nabla C$. The outcome space is Ω, and the contracts are determined by a payoff function $\phi : \Omega \to \mathbb{R}^n$ such that $\Pi \doteq \varphi(\mathbb{R}^n) = \mathsf{conv}(\phi(\Omega))$. That is, the derivative space Π of C (the "instantaneous prices") must be the convex hull of the payoffs.

A trader purchasing shares at the current prices $\pi \in \mathbb{R}^n$ pays $C(\varphi^{-1}(\pi) + r) - C(\varphi^{-1}(\pi))$ for the bundle of contracts $r \in \mathbb{R}^n$. Note that our dependence solely on π limits our model slightly, since in general the share space (domain of C) may contain more information than the current prices (cf. [1]). The bundle r is determined by an agent's *demand function* $d(C, \pi)$ which specifies the bundle to buy given the price π and the cost function C.

Our market dynamics are the following. The market maker posts the current price π_t, and at each time $t = 1 \dots T$, a trader is chosen with demand function d drawn i.i.d. from some demand distribution $\mathcal{D}$. Intuitively, these demands are parameterized by latent variables such as the agent's belief $p \in \Delta_\Omega$ and total wealth W. The price is then updated to

$$\pi_{t+1} = \varphi(\varphi^{-1}(\pi_t) + d(C, \pi_t)). \tag{6.1}$$

After update T, the outcome is revealed and payout $\phi(\omega)_i$ is given for each contract $i \in \{1, \dots, n\}$.

6.3 Stationarity and equilibrium

We first would like to relate our stochastic model (6.1) to the standard notion of market equilibrium from the Economics literature, which we call the Walrasian equilibrium to avoid confusion. Here prices are fixed, and the equilibrium price is one that clears the market, meaning that the sum of the demands r is $0 \in \mathbb{R}^n$. In fact, we will show that the stationary point of our process approaches the Walrasian equilibrium point as the liquidity of the market approaches infinity.

First, we must add a liquidity parameter to our market. Following the LMSR (the cost function $C(s) = b \ln \sum_i e^{s_i/b}$), we define

$$C_b(s) \doteq b\, C(s/b). \tag{6.2}$$

This transformation of a convex function is called a *perspective function* and is known to preserve convexity [25]. Observe that $\varphi_b(s) \doteq \nabla C_b(s) = \nabla C(s/b) = \varphi(s/b)$, meaning that the price under C_b at s is the same as the price under C at s/b. As with the LMSR, we call b the *liquidity parameter*; this terminology is justified by noting that one definition of liquidity, $1/\lambda_{\max}\nabla^2 C_b(s) = b/\lambda_{\max}\nabla^2 C(s/b)$ (cf. [1]). In the following, we will consider the limit as $b \to \infty$.

Second, in order to connect to the Walrasian equilibrium, we need a notion of a *fixed-price* demand function: if a trader has demand $d(C, \cdot)$ given C, what would the same trader's demand be under a market where prices are fixed and do not "change" during a trade? For the sake of generality, we restrict our allowable demand functions to the ones for which the limit

$$d(F, \pi) \doteq \lim_{b \to \infty} d(C_b, \pi) \tag{6.3}$$

exists; this demand $d(F, \cdot)$ will be the corresponding *fixed-price demand* for d. We now define the Walrasion equilibrium point π^*, which is simply the price at which the market clears when traders have demands distributed by $\mathcal{D}$. Formally, this is the following condition:[1]

$$\int_{\mathcal{D}} d(F, \pi^*)\, d\mathcal{D}(d) = 0 \tag{6.4}$$

[1]Here and throughout we ignore technical issues of uniqueness. One may simply restrict to the class of demands for which uniqueness is satisfied.

Note that $0 \in \mathbb{R}^n$; the demand for each contract should be balanced.

The stationary point of our stochastic process, on the other hand, is the price π_b^s for which the expected price fluctuation is 0. Formally, we have

$$\mathbb{E}_{d\sim\mathcal{D}}[\Delta(\pi_b^s, d(C_b, \pi_b^s))] = 0, \tag{6.5}$$

where $\Delta(\pi, d) \doteq \varphi(\varphi^{-1}(\pi) + d) - \pi$ is the price fluctuation. We now consider the limit of our stochastic process as the market liquidity approaches ∞.

Theorem 6.1. *Let C be a strictly convex and α-smooth[2] cost function, and assume that $\frac{\partial}{\partial b} d(C_b, \pi) = o(1/b)$ uniformly in π and all $d \in \mathcal{D}$. If furthermore the limit (6.3) is uniform in π and d, then $\lim_{b\to\infty} \pi_b^s = \pi^*$.*

Proof. Note that by the stationarity condition (6.5) we may define π^* and π_b^s to be the roots of the following "excess demand" functions, respectively:

$$Z(\pi) \doteq \int_{\mathcal{D}} d(F, \pi)\, d\mathcal{D}(d), \qquad Z_b^s(\pi) \doteq b\mathbb{E}_{d\sim\mathcal{D}}[\Delta(\pi, d(C_b, \pi))],$$

where we scale the latter by b so that Z_b^s does not limit to the zero function.

Let $s = \varphi^{-1}(\pi)$ be the current share vector. Then we have

$$\begin{aligned}
\lim_{b\to\infty} b\Delta(\pi, d(C_b, \pi)) &= \lim_{b\to\infty} b\left(\varphi\left(\varphi^{-1}(\pi) + d(C_b, \pi)/b\right) - \pi\right) \\
&= \lim_{a\to 0} \frac{\varphi\left(s + a\, d(C_{1/a}, \pi)\right) - \pi}{a} \\
&= \lim_{a\to 0} \nabla\varphi\left(s + a\, d(C_{1/a}, \pi)\right)\left(d(C_{1/a}, \pi) + a\tfrac{\partial}{\partial a} d(C_{1/a}, \pi)\right) \\
&= \lim_{b\to\infty} \nabla\varphi\left(s + \tfrac{1}{b}\, d(C_b, \pi)\right)\left(d(C_b, \pi) + \tfrac{1}{b}\tfrac{\partial}{\partial b} d(C_b, \pi)(-b^2)\right) \\
&= \lim_{b\to\infty} \nabla^2 C(s)\, d(C_b, \pi) \;=\; \nabla^2 C(s)\, d(F, \pi),
\end{aligned}$$

where we apply L'Hopital's rule for the third equality. Crucially, the above limit is uniform with respect to both $d \in \mathcal{D}$ and $\pi \in \Pi$; uniformity in d is by assumption, and uniformity in π follows from α-smoothness of C, since C is dominated by a quadratic. Since the limit is uniform with respect to $\mathcal{D}$, we now have

$$\begin{aligned}
\lim_{b\to\infty} Z_b^s(\pi) = \lim_{b\to\infty} b\mathbb{E}_{d\sim\mathcal{D}}[\Delta(\pi, d(C_b, \pi))] &= \mathbb{E}_{d\sim\mathcal{D}}\left[\lim_{b\to\infty} b\Delta(\pi, d(C_b, \pi))\right] \\
&= \nabla^2 C(s)\mathbb{E}_{d\sim\mathcal{D}}[d(F, \pi)] = \nabla^2 C(s)\, Z(\pi).
\end{aligned}$$

As $\nabla^2 C(s)$ is positive definite by assumption on C, we can conclude that $\lim_{b\to\infty} Z_b^s$ and Z share the same zeroes. Since Z has compact domain and is assumed continuous with a unique zero π^*, for each $\epsilon \in (0, \epsilon_{max})$ there must be some $\delta > 0$ s.t. $|Z(\pi)| > \epsilon$ for all π s.t.

[2] C is α-smooth if $\lambda_{\max}\nabla^2 C \leq \alpha$

$\|\pi - \pi^*\| > \delta$ (otherwise there would be a sequence of $\pi_n \to \pi'$ s.t. $f(\pi') = 0$ but $\pi' \neq \pi^*$). By uniform convergence there must be a $B > 0$ s.t. for all $b > B$ we have $\|Z_b^s - Z\|_\infty < \epsilon/2$. In particular, for π s.t. $\|\pi - \pi^*\| > \delta$, $|Z_b^s(\pi)| > \epsilon/2$. Thus, the corresponding zeros π_b^s must be within δ of π^*. Hence $\lim_{b\to\infty} \pi_b^s = \pi^*$.[3] □

6.3.1 Utility-based demands

Maximum Expected Utility (MEU) demand functions are a particular kind of demand function derived by assuming a trader has some belief $p \in \Delta^n$ over the outcomes in Ω, some wealth $W \geq 0$, and a monotonically increasing utility function of money $u : \mathbb{R} \to \mathbb{R}$. If such a trader buys a bundle r of contracts from a market maker with cost function C and price π, her wealth after ω occurs is $\Upsilon_\omega(C, W, \pi, r) \doteq W + \phi(\omega) \cdot r - [C(\varphi^{-1}(\pi) + r) - C(\varphi^{-1}(\pi))]$. We ensure traders do not go into debt by requiring that traders only make demands such that this final wealth is nonnegative: $\forall\omega\ \Upsilon_\omega(C, \pi, r) \geq 0$. The set of debt-free bundles for wealth W and market C at price π is denoted $S(C, W, \pi) \doteq \{r \in \mathbb{R}^n \ :\ \min_\omega \Upsilon_\omega(C, W, \pi, r) \geq 0\}$.

A *continuous MEU demand function* $d^u_{W,p}(C, \pi)$ is then just the demand that maximizes a trader's expected utility subject to the debt-free constraint. That is,

$$d^u_{W,p}(C, \pi) \doteq \operatorname*{argmax}_{r \in S(C,W,\pi)} \mathbb{E}_{\omega\sim p} \left[u \left(\Upsilon_\omega(C, W, \pi, r)\right)\right]. \tag{6.6}$$

We also define a *fixed-price MEU demand function* $d^u_{W,p}(F, \pi)$ similarly, where $\Upsilon_\omega(F, W, \pi, r) \doteq W + \phi(\omega) \cdot r - \pi \cdot r$ and $S(F, W, \pi) \doteq \{r \in \mathbb{R}^n \ :\ \min_\omega \Upsilon_\omega(F, W, \pi, r) \geq 0\}$ are the fixed price analogues to the continuously priced versions above. Using the notation $bS \doteq \{b\, r \mid r \in S\}$, the following relationships between the continuous and fixed price versions of Υ, S_W, and the expected utility are a consequence of the convexity of C. Their main purpose is to highlight the relationship between wealth and liquidity in MEU demands. In particular, they show that scaling up of liquidity is equivalent to a scaling down of wealth and that the continuously priced constraints and wealth functions monotonically approach the fixed priced versions.

Lemma 6.2. *For any strictly convex cost function C, wealth $W > 0$, price π, demand r, and liquidity parameter $b > 0$ the following properties hold:*

1. $\Upsilon_\omega(C_b, W, \pi, r) = b\, \Upsilon_\omega(C, W/b, \pi, r/b)$;
2. $S(C_b, W, \pi) = b\, S(C, W/b, \pi)$;
3. $S(C, W, \pi)$ *is convex for all* C;
4. $S(C, W, \pi) \subseteq S(C_b, W, \pi) \subseteq S(F, W, \pi)$ *for all* $b \geq 1$.
5. *For monotone utilities* u, $\mathop{\mathbb{E}}_{\omega\sim p} \left[u \left(\Upsilon_\omega(F, W, \pi, r)\right)\right] \geq \mathop{\mathbb{E}}_{\omega\sim p} \left[u \left(\Upsilon_\omega(C, W, \pi, r)\right)\right]$.

[3] We thank Avraham Ruderman for a helpful discussion regarding this proof.

Proof. Property (1) follows from a simple computation:

$$\begin{aligned}\Upsilon_\omega(C_b, W, \pi, r) &= W + \phi(\omega)\cdot r - b\,C(\varphi^{-1}(\pi) + r/b) + b\,C(\varphi^{-1}(\pi)) \\ &= b\left(W/b + \phi(\omega)\cdot(r/b) - C(\varphi^{-1}(\pi) + r/b) + C(\varphi^{-1}(\pi))\right),\end{aligned}$$

which equals $b\,\Upsilon_\omega(C, W/b, \pi, r/b)$ by definition. We now can see property (2) as well:

$$\begin{aligned}S(C_b, W, \pi) &= \{r \;:\; \min_\omega b\,\Upsilon_\omega(C, W/b, \pi, r/b) \geq 0\} \\ &= \{b\,r \;:\; \min_\omega \Upsilon_\omega(C, W/b, \pi, r) \geq 0\}.\end{aligned}$$

For (3), define $f_{C,s,\omega}(r) = C(s+r) - C(s) - \phi(\omega)\cdot r$, which is the ex-post cost of purchasing bundle r. As C is convex, and $f_{C,s,\omega}$ is a shifted and translated version of C plus a linear term, $f_{C,s,\omega}$ is convex also. The constraint $\Upsilon_\omega(C, W, \pi, r) \geq 0$ then translates to $f_{C,s,\omega}(r) \leq W$, and thus the set of r which satisfy the constraint is convex as a sublevel set of a convex function. Now $S(C, W, \pi)$ is convex as an intersection of convex sets, proving (3).

For (4) suppose r satisfies $f_{C,s,\omega}(r) \leq W$. Note that $f_{C,s,\omega}(0) = 0$ always. Then by convexity we have for $f := f_{C,s,\omega}$ we have $f(r/b) = f\left(\frac{1}{b}r + \frac{b-1}{b}0\right) \leq \frac{1}{b}f(r) + \frac{b-1}{b}0 \leq W/b$, which implies $S(C, W, \pi) \subseteq S(C_b, W, \pi)$ when considering (3). To complete (4) note that $f_{C,s,\omega}$ dominates $f_{F,s} : r \mapsto (\varphi(s) - \phi(\omega))\cdot r$ by convexity of C: $C(s+r) - C(s) \geq \nabla C(s)\cdot r$.

Finally, proof of (5) is obtained by noting that the convexity of C means that $C(\varphi^{-1}(\pi) + r) - C(\varphi^{-1}(\pi)) \geq \nabla C(\varphi^{-1}(\pi))\cdot r = \pi\cdot r$ and exploting the monotonicty of u. □

Lemma 6.2 shows us that MEU demands have a lot of structure, and in particular, properties (4) and (5) suggest that they may satisfy the conditions of Theorem 6.1; we leave this as an open question for future work. Another interesting aspect of Lemma 6.2 is the relationship between markets with cost function C_b and wealths W and markets with cost function C and wealths W/b – indeed, properties (1) and (2) suggest that the liquidity limit should in some sense be equivalent to a wealth limit, in that increasing liquidity by a factor b should yield similar dynamics to decreasing the wealths by b. This would relate our model to that of Othman and Sandholm [81], where the authors essentially show a wealth-limit version of Theorem 6.1 for a binary-outcome market where traders have linear utilities (a special case of (6.6)). We leave this precise connection for future work.

6.4 Market making as mirror descent

We now explore the surprising relationship between our stochastic price update and standard stochastic optimization techniques. In particular, we will relate our model to a stochastic *mirror descent* of the form

$$x_{t+1} = \operatorname*{argmin}_{x\in\mathbb{R}}\{\eta\, x\cdot\nabla F(x_t;\xi) + D_R(x, x_t)\}, \tag{6.7}$$

where at each step $\xi \sim \Xi$ are i.i.d. and R is some strictly convex function. We will refer to an algorithm of the form (6.7) a *stochastic mirror descent* of $f(x) \doteq \mathbb{E}_{\xi\sim\Xi}[F(x;\xi)]$.

Theorem 6.3. *If for all $d \in \mathcal{D}$ we have some $F(\cdot\,; d) : \mathbb{R}^n \to \mathbb{R}^n$ such that $d(R^*, \pi) = -\nabla F(\pi; d)$, then the stochastic update of our model (6.1) is exactly a stochastic mirror descent of $f(\pi) = \mathbb{E}_{d\sim\mathcal{D}}[F(\pi; d)]$.*

Proof. By standard arguments, the mirror descent update (6.7) can be rewritten as

$$x_{t+1} = \nabla R^*(\nabla R(x_t) - \nabla F(x_t; \xi)),$$

where R^* is the conjugate dual of R. Take $R = C^*$, and let $\xi = d \sim \mathcal{D}$. By assumption, we have $\nabla F(x; d) = -d(R^*, x) = -d(C, x)$ for all d. As $\nabla R^* = \nabla C = \varphi$, we have $\varphi^{-1} = (\nabla R^*)^{-1} = \nabla R$ by duality, and thus our update becomes $x_{t+1} = \varphi\left(\varphi^{-1}(x_t) + d(C, x_t)\right)$, which exactly matches the stochastic update of our model (6.1). □

As an example, consider Kelly betters, which correspond to fixed-price demands $d(C, \pi) \doteq d^{\log}_{W,p}(F, \pi)$ with utility $u(x) = \log x$ as defined in (6.3). A simple calculation shows that our update becomes

$$\pi_{t+1} = \varphi\left(\varphi^{-1}(\pi_t) + \frac{W}{\pi}\,\frac{p-\pi}{1-\pi}\right), \tag{6.8}$$

where W and p are drawn (independently) from $\mathcal{P}$ and $\mathcal{W}$.

Corollary 6.4. *The stochastic update for fixed-price Kelly betters (6.8) is exactly a stochastic mirror descent of $f(\pi) = \overline{W} \cdot \mathrm{KL}(\overline{p}, \pi)$, where $\overline{p}$ and $\overline{W}$ are the means of $\mathcal{P}$ and $\mathcal{W}$, respectively.*

Proof. We take $F(x; d^{\log}_{W,p}) = W \cdot (\mathrm{KL}(p, x) + H(p))$. Then

$$\nabla F(x; d^{\log}_{W,p}) = W\left(\frac{-p}{x} + \frac{p-1}{1-x}\right) = -\frac{W}{x}\,\frac{p-x}{1-x} = -d^{\log}_{W,p}(F, x).$$

Hence, by Theorem 6.3 our update is a stochastic mirror descent of:

$$\begin{aligned} f(x) &\doteq \mathbb{E}[F(x; d^{\log}_{W,p})] \\ &= \mathbb{E}[Wp\log x + W(1-p)\log(1-x)] \\ &= \overline{W} \cdot (\mathrm{KL}(\overline{p}, x) + H(\overline{p}))\,, \end{aligned}$$

which of course is equivalent to $\overline{W} \cdot \mathrm{KL}(\overline{p}, x)$ as the entropy term does not depend on x. □

Note that while this last result is quite compelling, we have mixed fixed-price demands with a continuous-price market model — see §6.3.1. One could interpret this combination as a model in which the market maker can only adjust the prices *after* a trade, according to a fixed convex cost function C. This of course differs from the standard model, which adjusts the price *continuously* during a trade.

6.4.1 Leveraging existing learning results

Theorem 6.3 not only identifies a fascinating connection between machine learning and our stochastic prediction market model, but it also allows us to use powerful existing techniques to make broad conclusions about the behavior of our model. Consider the following result:

Proposition 6.5 ([42]). *If $\|\nabla F(\pi; p)\|^2 \leq G^2$ for all p, π, and R is σ-strongly convex, then with probability $1 - \delta$,*

$$f(\overline{\pi}_T) \leq \min_{\pi} f(\pi) + \left(\frac{D^2}{\eta T} + \frac{G^2 \eta}{2\sigma}\right)\left(1 + 4\sqrt{\log \frac{1}{\delta}}\right).$$

In our context, Proposition 6.5 says that the *average* of the prices will be a very good estimate of the minimizer of f, which as suggested by happens to be the underlying mean belief $\overline{p}$ of the traders! Moreover, as the Kelly demands are linear in both p and W, it is easy to see from Theorem 6.1 that $\overline{p}$ is also the stationary point and the Walrasian equilibrium point (the latter was also shown by Wolfers and Zitzewitz [102]). On the other hand, as we demonstrate next, it is not hard to come up with an example where the *instantaneous* price π_t is quite far from the equilibrium at any given time period.

Before moving to empirical work, we make one final point. The above relationship between our stochastic market model and mirror descent sheds light on an important question: how might an automated market maker adjust the liquidity so that the market actually converges to the mean of the traders' beliefs? The learning parameter η can be thought of as the inverse of the liquidity, and as such, Proposition 6.5 suggests that increasing the liquidity as $\sqrt{t}$ may cause the mean price to converge to the mean belief (assuming a fixed underlying belief distribution).

6.5 Empirical work

Example: biased coin

Consider a classic Bayesian setting where a coin has unknown bias Pr[heads] $= q$, and traders have a prior $\beta(\alpha, \alpha)$ over q (i.e., traders are α-confident that the coin is fair). Now suppose each trader independently observes n flips from the coin, and updates her belief; upon seeing k heads, a trader would have posterior $\beta(\alpha + k, \alpha + n - k)$.

When presented with a prediction market with contracts for a single toss of the coin, where and contract 0 pays \$1 for tails and contract 1 pays \$1 for heads, a trader would purchase contracts as if according to the mean of their posterior. Hence, the belief distribution $\mathcal{P}$ of the market assigns weight $\mathcal{P}(p) = \binom{n}{k} q^k (1-q)^{n-k}$ to belief $p = (\alpha + k)/(2\alpha + n)$, yielding a biased mean belief of $(\alpha + nq)/(2\alpha + n)$.

We show a typical simulation of this market in Figure 6.1, where traders behave as Kelly betters in the fixed-price LMSR. Clearly, after almost every trade, the market price is

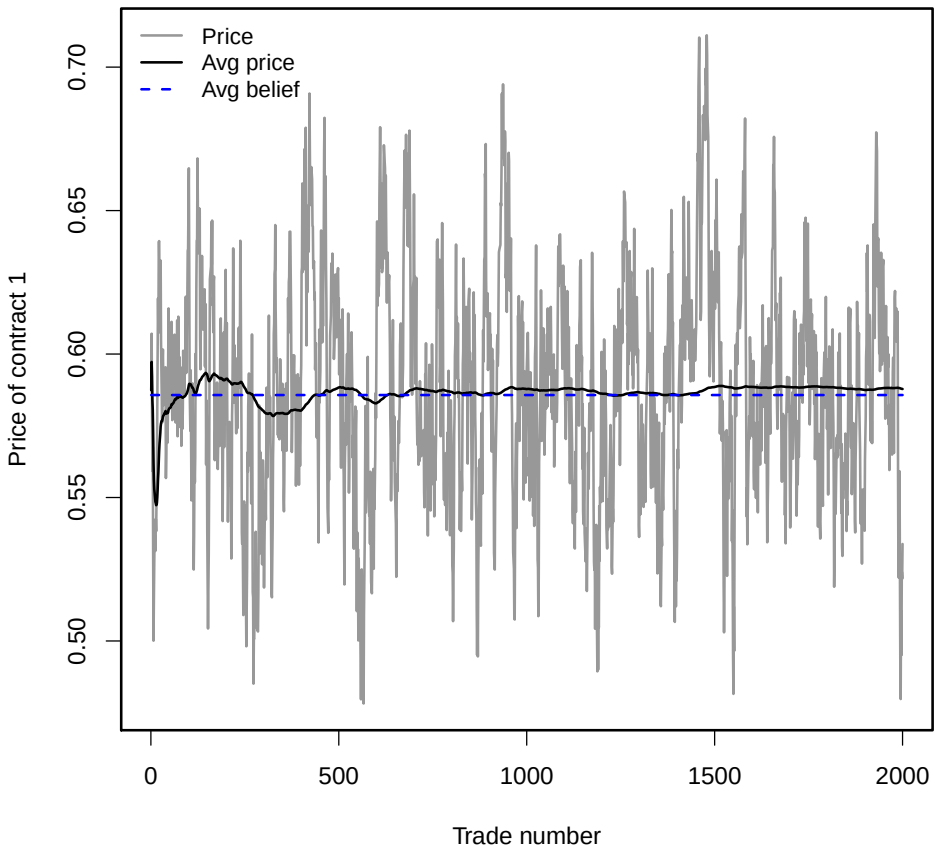


Figure 6.1: Price movement for Kelly betters with binomial($q = 0.6$, $n = 6$, $\alpha = 0.5$) beliefs in the LMSR market with liquidity $b = 10$.

quite far from the equilibrium/stationary point, and hence the classical supply and demand analysis of this market yields a poor description of the actual behavior, and in particular, of the predictive quality of the price at any given time. However, the mean price is consistently close to the mean belief of the traders, which in turn is quite close to the true parameter q.

Election Survey Data

We now compare the quality of the running average price versus the instantaneous price as a predictor of the mean belief of a market. We do so by simulating a market maker interacting with traders with unit wealth, log utility, and beliefs drawn from a fixed distribution. The

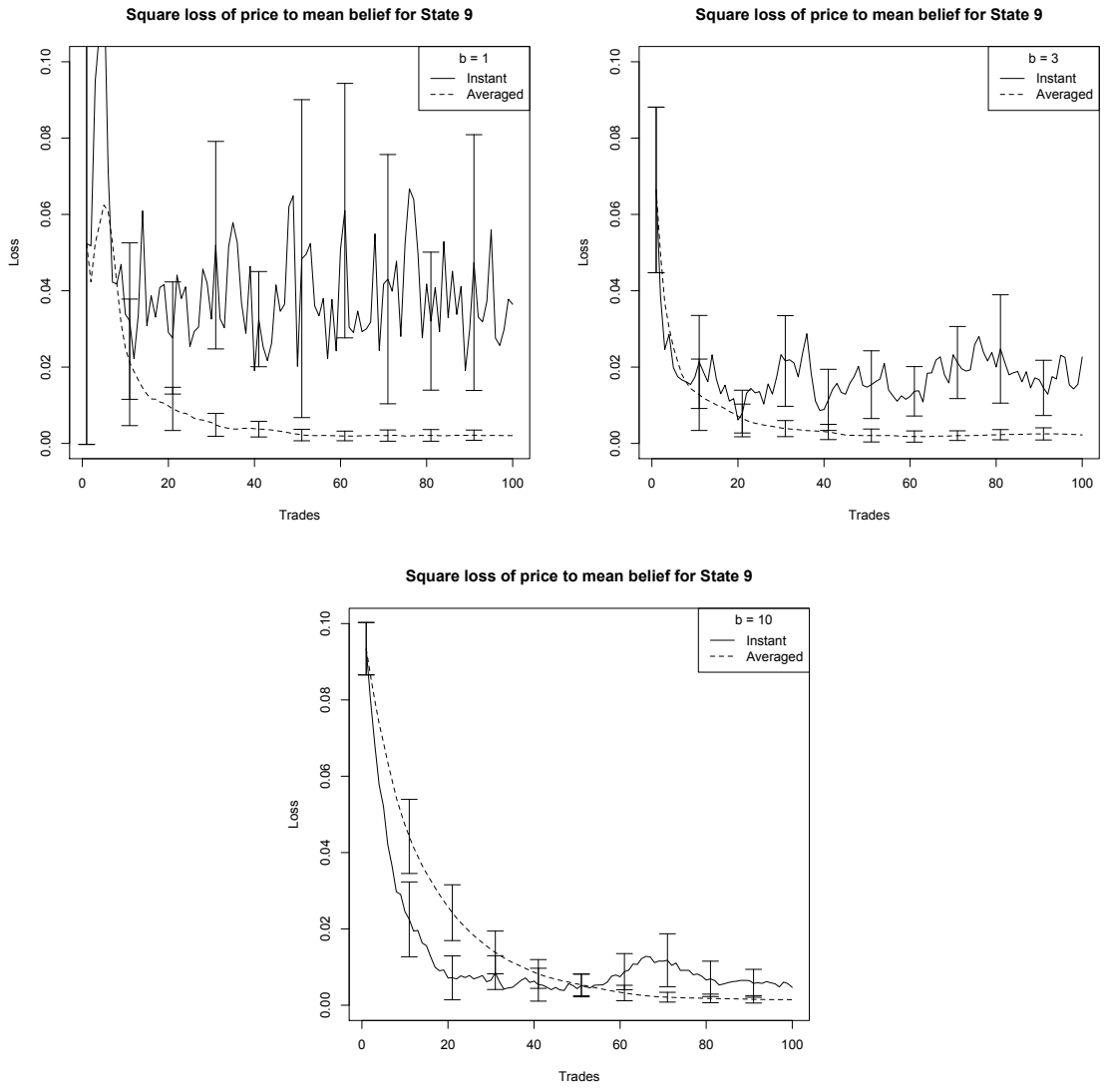


Figure 6.2: Mean square loss of average and instantaneous prices relative to the mean belief of 0.26 over 20 simulations for State 9 for $b = 1$ (left), $b = 3$ (middle), and $b = 10$ (right). Bars show standard deviation.

belief distributions are derived from the Princeton election survey data [99]. For each of the 50 US states, participants in the survey were asked to estimate the probability that one of two possible candidates were going to win that state.[4] We use these 50 sets of estimates as 50 different empirical distributions from which to draw trader beliefs.

A simulation is configured by choosing one of the 50 empirical belief distributions S, a

[4] The original dataset contains conjunctions of wins as well as conditional statements but we only use the single variable results of the survey.

market liquidity parameter b to define the LMSR cost function $C(s) = b \ln \sum_i e^{s_i/b}$, and an initial market position vector of $(0, 0)$ – that is, no contracts for either outcome. A configured simulation is run for T trades. At each trade, a belief p is drawn from S uniformly and with replacement. This belief is used to determine the demand of the trader relative to the current market pricing. The trader purchase a bundle of contracts according to its demand and the market moves its position and price accordingly. The complete price path π_t for $t = 1, \ldots, T$ of the market is recorded as well as a running average price $\bar{\pi}_t \doteq \frac{1}{t} \sum_{i=1}^{t} \pi_t$ for $t = 1 \ldots, T$. For each of the 50 empirical belief distributions we configured 9 markets with $b \in \{1, 2, 3, 5, 10, 15, 20, 30, 50\}$ and ran 20 independent simulations of $T = 100$ trades. We present a portion of the results for the empirical distributions for states 9 and 11. States 9 and 11 have, respectively, sample sizes of 2,717 and 2,709; means 0.26 and 0.9; and variances 0.04 and 0.02. These are chosen as being representative of the rest of the simulation results: State 9 with mean off-center and a spread of beliefs (high uncertainty) and State 11 with highly concentrated beliefs around a single outcome (low uncertainty).

The results are summarised in Figures 6.2, 6.3, and 6.4. The first show the square loss of the average and instaneous prices relative to the mean belief for high uncertainty State 9 for $b = 1, 3, 10$. Clearly, the average price is a much more reliable estimator of the mean belief for low liquidity ($b = 1$) and is only outperformed by the instaneous price for higher liquidity ($b = 10$), but then only early in trading. Similar plots for State 11 are shown in Figure 6.3 where the advantage of using the average price is significantly diminished.

Figure 6.4 shows the improvement the average price has over the instantaneous price in square loss relative to the mean belief for all liquidity settings and highlights that average prices work better in low liquidity settings, consistent with the theory. Similar trends were observed for all the other States, depending on whether they had high uncertainty – in which case average price was a much better estimator – or low uncertainty – in which case instanteous price was better.

6.6 Conclusion and future work

As noted in §6.3.1, there are several open questions with regard to maximum expected utility demands and Theorem 6.1, as well as the relationship between trader wealth and market liquidity. It would also be interesting to have a application of Theorem 6.3 to a continuous-price model, which yields a natural minimization as in Corollary 6.4. The equivalence to mirror decent stablished in Theorem 6.3 may also lead to a better understanding of the optimal manner in which a automated prediction market ought to increase liquidity so as to maximise efficiency.

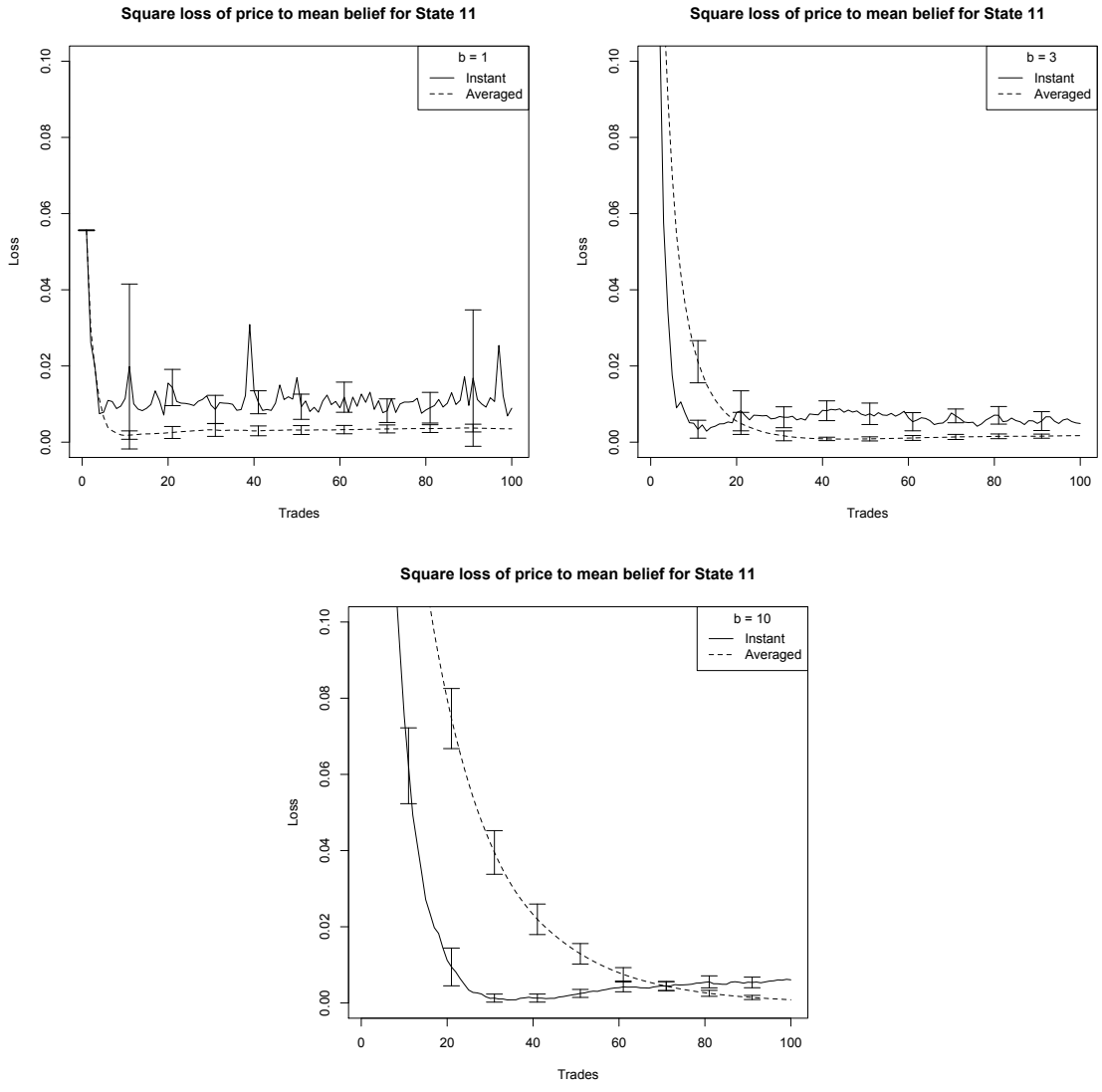


Figure 6.3: Mean square loss of average and instantaneous prices relative to the mean belief of 0.9 over 20 simulations for State 11 for $b = 1$ (left), $b = 3$ (middle), and $b = 10$ (right). Bars show standard deviation.

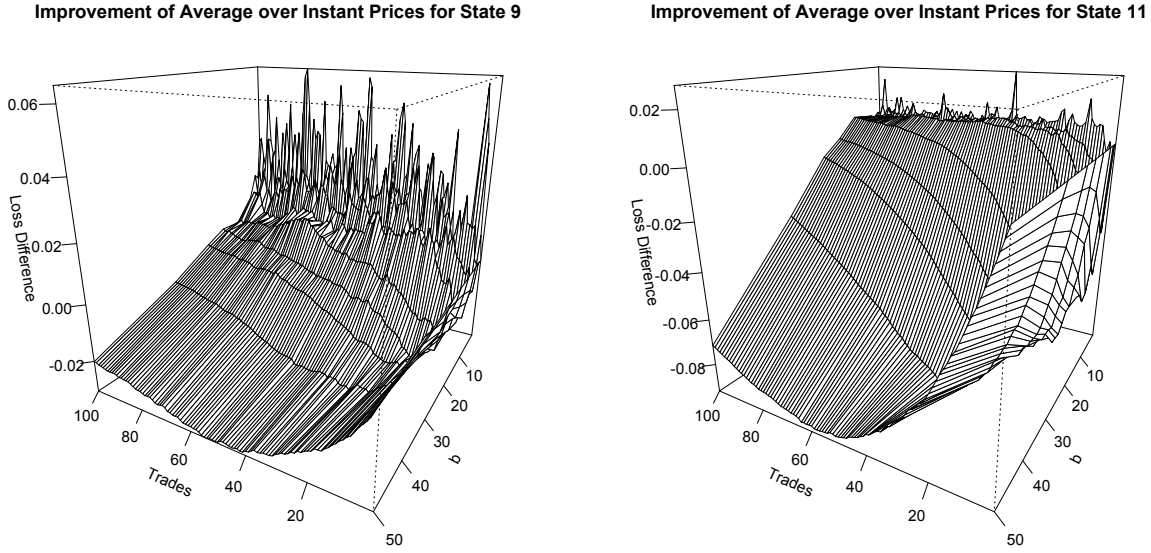


Figure 6.4: An overview of the results for States 9 (left) and 11 (right). For each trade and choice of b, the vertical value shows the improvement of the average price over the instantaneous price as measure by square loss relative to the mean.

Bibliography

[1] J. Abernethy, Y. Chen, and J. Wortman Vaughan. An optimization-based framework for automated market-making. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 297–306, 2011.

[2] J. Abernethy and R. Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the 25th Conference on Learning Theory*, 2012.

[3] Jacob Abernethy, Yiling Chen, and Jennifer Wortman Vaughan. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation*, 1(2):12, 2013.

[4] Jacob D. Abernethy and Rafael M. Frongillo. A collaborative mechanism for crowdsourcing prediction problems. In *Advances in Neural Information Processing Systems 24*, pages 2600–2608, 2011.

[5] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2007.

[6] Shun'ichi Amari. *Differential-geometrical methods in statistics*. Springer-Verlag, Berlin; New York, 1985.

[7] A. Archer and R. Kleinberg. Truthful germs are contagious: a local to global characterization of truthfulness. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 21–30, 2008.

[8] I. Ashlagi, M. Braverman, A. Hassidim, and D. Monderer. Monotonicity and implementability. *Econometrica*, 78(5):1749–1772, 2010.

[9] F. Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

[10] F. Aurenhammer. Recognising polytopical cell complexes and constructing projection polyhedra. *Journal of Symbolic Computation*, 3(3):249–255, June 1987.

[11] Franz Aurenhammer. A criterion for the affine equivalence of cell complexes in r d and convex polyhedra in r d+1. *Discrete & Computational Geometry*, 2(1):49–64, December 1987.

[12] Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[13] M. Babaioff, L. Blumrosen, N. S. Lambert, and O. Reingold. Only valuable experts can be valued. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 221–222, 2011.

[14] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88, 2009.

[15] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, July 2005.

[16] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[17] Adrian Barbu and Nathan Lay. An introduction to artificial prediction markets for classification. *Journal of Machine Learning Research*, 13:2177–2204, 2012.

[18] Ole E Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley, Chichester, 1978.

[19] J. E. Berg, R. Forsythe, F. D. Nelson, and T. A. Rietz. Results from a dozen years of election futures markets research. In C. A. Plott and V. Smith, editors, *Handbook of Experimental Economic Results*, volume 1, pages 742–751. Elsevier, 2008.

[20] A. Berger, R. Mller, and S. Naeemi. Characterizing incentive compatibility for convex valuations. *Algorithmic Game Theory*, pages 24–35, 2009.

[21] Andr Berger, Rudolf Mller, and Naeemi Seyed Hossein. Path-monotonicity and incentive compatibility. Technical report, Maastricht: METEOR, Maastricht Research School of Economics of Technology and Organization, 2010.

[22] Alina Beygelzimer, John Langford, and David M Pennock. Learning performance of prediction markets with kelly bettors. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, volume 3, pages 1317–1318, 2012.

[23] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman voronoi diagrams: Properties, algorithms and applications. *CoRR*, abs/0709.2196, 2007.

[24] Jonathan M. Borwein and Jon D. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*. Cambridge University Press, January 2010.

[25] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[26] G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):13, 1950.

[27] Yang Cai, Mohammad Mahdian, Aranyak Mehta, and Bo Waggoner. Designing markets for daily deals. *Preprint*, 2013.

[28] G. Carroll. When are local incentive constraints sufficient? *Econometrica*, 80(2):661–686, 2012.

[29] Y. Chen, L. Fortnow, N. Lambert, D. M Pennock, and J. Wortman. Complexity of combinatorial market makers. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 190–199, 2008.

[30] Y. Chen, L. Fortnow, E. Nikolova, and D.M. Pennock. Betting on permutations. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, pages 326–335, 2007.

[31] Y. Chen, I. Kash, M. Ruberry, and V. Shnayder. Decision markets with good incentives. *Internet and Network Economics*, pages 72–83, 2011.

[32] Y. Chen and I. A Kash. Information elicitation for decision making. *AAMAS*, 2011.

[33] Y. Chen and D.M. Pennock. A utility framework for bounded-loss market makers. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 49–56, 2007.

[34] Y. Chen and J.W. Vaughan. A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 189–198, 2010.

[35] K. S. Chung and J. C. Ely. Ex-post incentive compatible mechanism design. *URL http://www.kellogg.northwestern.edu/research/math/dps/1339.pdf. Working Paper*, 2002.

[36] J. Cid-Sueiro. Proper losses for learning from partial labels. In *Advances in Neural Information Processing Systems 25*, pages 1574–1582, 2012.

[37] Edward H. Clarke. Multipart pricing of public goods. *Public choice*, 11(1):17–33, 1971.

[38] T.M. Cover, J.A. Thomas, J. Wiley, et al. *Elements of information theory*, volume 6. Wiley Online Library, 1991.

[39] A. Philip Dawid. Coherent measures of discrepancy, uncertainty and dependence, with applications to bayesian predictive experimental design. Technical report, Technical Report 139, Department of Statistical Science, University College London. http://www.ucl.ac.uk/Stats/research/abs94.html, 1998.

[40] Shahar Dobzinski and Noam Nisan. Mechanisms for multi-unit auctions. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 346–351, 2007.

[41] Shahar Dobzinski and Noam Nisan. A modular approach to roberts theorem. *Algorithmic Game Theory*, pages 14–23, 2009.

[42] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. *Technical Report*, 2010.

[43] Fang Fang, Maxwell B. Stinchcombe, and Andrew B. Whinston. Proper scoring rules with arbitrary value functions. *Journal of Mathematical Economics*, 46(6):1200–1210, 2010.

[44] Jon Feldman and S. Muthukrishnan. Algorithmic methods for sponsored search advertising. In Zhen Liu and Cathy H. Xia, editors, *Performance Modeling and Engineering*, pages 91–122. Springer US, January 2008.

[45] Amos Fiat, Anna Karlin, Elias Koutsoupias, and Angelina Vidali. Approaching utopia: strong truthfulness and externality-resistant mechanisms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 221–230, 2013.

[46] R. Frongillo, N. Della Penna, and M. Reid. Interpreting prediction markets: a stochastic approach. In *Advances in Neural Information Processing Systems 25*, pages 3275–3283, 2012.

[47] Rafael M. Frongillo and Ian A. Kash. General truthfulness characterizations via convex analysis. *arXiv:1211.3043*, November 2012.

[48] Francis Galton. Vox populi (the wisdom of crowds). *Nature*, 75, 1907.

[49] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

[50] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[51] T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[52] Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

[53] Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631, 1973.

[54] P.D. Grnwald and A.P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

[55] J.Y. Halpern. *Reasoning about uncertainty*. MIT Press, 2003.

[56] R. Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5(1):107–119, 2003.

[57] Birgit Heydenreich, Rudolf Mller, Marc Uetz, and Rakesh V. Vohra. Characterization of revenue equivalence. *Econometrica*, 77(1):307–316, 2009.

[58] Nathanal Hyafil and Craig Boutilier. *Mechanism design with partial revelation*. University of Toronto, 2008.

[59] Aleksandr Davidovich Ioffe and V. M Tikhomirov. *Theory of extremal problems*. North-Holland Pub. Co. ; sole distributors for the U.S.A. and Canada, Elsevier North-Holland, Amsterdam; New York; New York, 1979.

[60] Rishabh Iyer and Jeff Bilmes. The lovsz-bregman divergence and connections to rank aggregation, clustering, and web ranking: Extended version. *Uncertainity in Artificial Intelligence*, 2013.

[61] Vijay Krishna and Eliot Maenner. Convex potentials with an application to mechanism design. *Econometrica*, 69(4):1113–1119, July 2001.

[62] Hang-Chin Lai and Lai-Jui Lin. The fenchel-moreau theorem for set functions. *Proceedings of the American Mathematical Society*, pages 85–90, 1988.

[63] N.S. Lambert. Elicitation and evaluation of statistical forecasts. *Preprint*, 2011.

[64] N.S. Lambert, J. Langford, J.W. Vaughan, Y. Chen, D. Reeves, Y. Shoham, and D.M. Pennock. An axiomatic characterization of wagering mechanisms. Technical report, Working paper, 2011.

[65] N.S. Lambert, D.M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

[66] N.S. Lambert and Y. Shoham. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 109–118, 2009.

[67] R. Lavi, A. Mualem, and N. Nisan. Two simplified proofs for roberts theorem. *Social Choice and Welfare*, 32(3):407–423, 2009.

[68] Ron Lavi, Ahuva Mu'Alem, and Noam Nisan. Towards a characterization of truthful combinatorial auctions. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*, pages 574–583, 2003.

[69] J. Ledyard, R. Hanson, and T. Ishikida. An experimental test of combinatorial information markets. *Journal of Economic Behavior & Organization*, 69(2):182–189, 2009.

[70] C. F. Manski. Interpreting the predictions of prediction markets. *Economics Letters*, 91(3):425–429, 2006.

[71] J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42(9):654, 1956.

[72] R. B. Myerson. Optimal auction design. *Mathematics of operations research*, pages 58–73, 1981.

[73] R. Mller, A. Perea, and S. Wolf. Weak monotonicity and bayes-nash incentive compatibility. *Games and Economic Behavior*, 61(2):344–358, 2007.

[74] Y. Narahari and S. Gujar. The mechanism design environment. *Tutorial*, 2012.

[75] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *arXiv preprint arXiv:1010.2731*, 2010.

[76] Frank Nielsen and Richard Nock. Entropies and cross-entropies of exponential families. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3621–3624, 2010.

[77] Noam Nisan. *Algorithmic game theory*. Cambridge University Press, 2007.

[78] Kent Osband. Optimal forecasting incentives. *Journal of Political Economy*, 97(5):1091–1112, October 1989.

[79] Kent Osband and Stefan Reichelstein. Information-eliciting compensation schemes. *Journal of Public Economics*, 27(1):107–115, June 1985.

[80] A. Othman and T. Sandholm. Decision rules and decision markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, volume 1, pages 625–632, 2010.

[81] A. Othman and T. Sandholm. When do markets with simple agents fail? In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 865–872, 2010.

[82] D.M. Pennock and R. Sami. Computational aspects of prediction markets. *Algorithmic Game Theory*, pages 651–674, 2007.

[83] Vladimir Rakocevic. On continuity of the moore-penrose and drazin inverses. *Matematiki Vesnik*, 49:163–172, 1997.

[84] M.D. Reid and R.C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 9999:2387–2422, 2010.

[85] Kevin Roberts. The characterization of implementable choice rules. *Aggregation and revelation of preferences*, 12(2):321–348, 1979.

[86] J. C. Rochet. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of Mathematical Economics*, 16(2):191–200, 1987.

[87] Jean-Charles Rochet and Lars A. Stole. The economics of multidimensional screening. *Econometric Society Monographs*, 35:150–197, 2003.

[88] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, October 2011.

[89] R.T. Rockafellar. *Convex analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, 1997.

[90] K. Rybnikov. Stresses and liftings of cell-complexes. *Discrete & Computational Geometry*, 21(4):481–517, June 1999.

[91] M. Saks and L. Yu. Weak monotonicity suffices for truthfulness on convex domains. In *Proceedings of the 6th ACM conference on Electronic commerce*, pages 286–293, 2005.

[92] L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, pages 783–801, 1971.

[93] M. J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989.

[94] A. Storkey. Machine learning markets. *Arxiv preprint arXiv:1106.4509*, 2011.

[95] Jean-Baptiste Hiriart Urruty and Claude Lemarchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[96] E. Vernet, R.C. Williamson, and M.D. Reid. Composite multiclass losses. *NIPS*, 2011.

[97] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

[98] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[99] G. Wang, S.R. Kulkarni, H.V. Poor, and D.N. Osherson. Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, 8(2):128, 2011.

[100] Wikipedia. Fenchel–moreau theorem, May 2013. Page Version ID: 509323438.

[101] J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of Economic Perspective*, 18(2):107–126, 2004.

[102] J. Wolfers and E. Zitzewitz. Interpreting prediction market prices as probabilities. Technical report, National Bureau of Economic Research, 2006.

[103] C. Zlinescu. *Convex Analysis in General Vector Spaces*. World Scientific Publishing Company, Incorporated, January 2002.